Novel Methods to Elucidate Core Classes in Multi-Dimensional Biomedical Data

by

Daniele Soria

Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy

School of Computer Science
The University of Nottingham
Nottingham, United Kingdom

March 2010

To my family with love;
to my friends with pride.

Abstract

Breast cancer, which is the most common cancer in women, is a complex disease characterised by multiple molecular alterations. Current routine clinical management relies on availability of robust clinical and pathologic prognostic and predictive factors, like the Nottingham Prognostic Index, to support decision making. Recent advances in high-throughput molecular technologies supported the evidence of a biologic heterogeneity of breast cancer.

This thesis is a multi-disciplinary work involving both computer scientists and molecular pathologists. It focuses on the development of advanced computational models for the classification of breast cancer into sub-types of the disease based on protein expression levels of selected markers. In a previous study conducted at the University of Nottingham, it has been suggested that immunohistochemical analysis may be used to identify distinct biological classes of breast cancer.

The objectives of this work were related both to the clinical and technical aspects. From a clinical point of view, the aim was to encourage a multiple techniques approach when dealing with classification and clustering. From a technical point of view, one of the goals was to verify the stability of groups obtained from different unsupervised clustering algorithms, applied to the same data, and to compare and combine the different solutions with the ones available from the previous study. These aims and objectives were considered in the attempt to fill a number of gaps in the body of knowledge. Several research questions were raised, including how to combine the results obtained by a multi-techniques approach for clustering and whether the medical decision making process could be moved in the direction of personalised healthcare.

An original framework to identify core representative classes in a dataset was developed and is described in this thesis. Using different clustering algorithms and several validity indices to explore the best number of groups to split the data, a set of classes may be defined by considering those points that remain stable across different clustering techniques. This set of representative classes may be then characterised resorting to usual statistical techniques and validated using supervised learning. Each step of this framework has been studied separately, resulting in different chapters of this thesis. The whole

approach has been successfully applied to a novel set of histone markers for breast cancer provided by the School of Pharmacy at the University of Nottingham. Although further tests are needed to validate and improve the proposed framework, these results make it a good candidate for being transferred to the real world of medical decision making.

Other contributions to knowledge may be extracted from this work. Firstly, six breast cancer subtypes have been identified, using consensus clustering, and characterised in terms of clinical outcome. Two of these classes were new in the literature. The second contribution is related to supervised learning. A novel method, based on the naive Bayes classifier, was developed to cope with the non-normality of covariates in many real world problems. This algorithm was validated over known data sets and compared with traditional approaches, obtaining better results in two examples.

All these contributions, and especially the novel framework may also have a clinical impact, as the overall medical care is gradually moving in the direction of a personalised one. By training a small number of doctors it may be possible for them to use the framework directly and find different sub-types of the disease they are investigating.

Acknowledgements

This thesis is dedicated to my whole family in Italy, and in particular to my mum Orietta, my dad Marco and my brother Alex for their unconditional love and unfailing encouragements throughout my recent and previous studies.

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Jon Garibaldi for giving me the opportunity of studying for my Ph.D. in Nottingham and for providing continuous support and guidance during my work. I am extremely grateful for his valuable comments and for the help he has afforded throughout this research programme.

I would also like to acknowledge Prof. Elia Biganzoli and Dr. Patrizia Boracchi for suggesting my name to my supervisor three years ago and encouraging me to follow this career. Together with Dr. Federico Ambrogi they have always been supportive and ready to help me; I am also extremely grateful for their continuous advice during my visiting periods in Milan.

My sincere gratitude also goes to Prof. Ian O. Ellis, Dr. Andy Green and all the people involved in the breast cancer research with which I have collaborated, not only for providing the data I have used, but also for their assistance with all biological issues and for making a big effort to understand the 'language' of a computer scientist. Many thanks also to Prof. David M. Heery and his Ph.D. student Magdy Korashy Abdel Fatah for providing me the data for validation of the last piece of this work.

This thesis would have never come to life, if not for the support of some amazing people. I would like to start by acknowledging all the fellows of the IMA and ASAP Research Groups in the School of Computer Science and those involved in the BIOPTRAIN programme around Europe.

In particular, I express my extreme gratitude to Dr. Julie Greensmith for being always available for a coffee and a chat whenever I needed them and for being involved in the 'R users' group.

Many thanks to my housemates Dr. German Terrazas Angulo, Paweł Widera and Pedro Leite Rocha and to other fellows Bob Oates, Elkin Castro, Enrico Glaab, Dr. Fran

Romero-Campero, James Smaldon, Dr. Jaume Bacardit, Juan Pedro Castro, Linda Fiaschi, Dr. Marcin Jaroszewski, Maria Franco, Sven Groenemeyer and Tiago Pais who have always been available for lunch or after-work drinks and for making my social life in Nottingham extremely enjoyable.

A special acknowledgement goes to Chiara Giovanelli for her friendship and encouragements throughout these years of life abroad and for her hospitality during my weekends in London.

I cannot forget some amazing people in Milan, who have always shown me their friendship and support no matter the physical distance between us. Many thanks to Silvia Monaco, Marta Bastia, Marta Giorgetti and Stefano Patti, Arvid Perego, Mara Picchetti, Francesco Rusconi, Marcella Nicolini, Mariella D'Alessio and Rocco Romano, Roberto Suardi, Antonella D'Alessio and Tiziano Peyla, Chiara Romano, Simone Mortara, Micol Metzinger, Cristiana Gorla and Alan Sorani and all of those I forgot to mention, but you know who you are!

Finally, I would like to thank once again my parents Orietta and Marco, my brother Alex and his wife Francesca for their continuous support and for enabling me to pursue my dreams, whatever they may be.

This work was supported by Marie Curie Action MEST-CT-2004-7597 under the Sixth Framework Programme (FP6) of the European Community.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Worldwide, cancer has become a major issue for human health. The classification of cancer patients is of great importance for its prognosis. In the last few years, many unsupervised and supervised algorithms have been proposed for this task and modern machine learning techniques are progressively being used by biologists to obtain proper tumour information from databases. This thesis investigates different classification algorithms for breast cancer data and proposes a step-by-step guideline to identify core classes in data sets in an attempt to produce an automated model for breast cancer classification which may be useful for future patients presenting at any hospital. This chapter provides the background and motivation for this research and introduces its aims and objectives. Later, an outline of the thesis organisation is reported.

1.1 Background and Motivation

The World Health Organization's Global Burden of Disease statistics identified cancer as the second largest global cause of death, after cardiovascular disease [23]. Cancer is the fastest growing segment of the disease burden; global cancer deaths are projected to increase from 7.1 million in 2002 to 11.5 million in 2030 [133]. Among them, breast cancer is the second most common type of cancer after lung cancer (10.4% of all cancer incidence, both sexes counted) [82] and the fifth most common cause of cancer death [81].

Breast cancer is a common disease which affects mostly (but not only) women. The ability to accurately identify the malignancy is crucial for prognosis and preparation of effective treatment. Breast cancer is usually, but not always, primarily classified by its histological appearance [181]. The first symptom, or subjective sign, of breast cancer is typically a lump that feels different from the surrounding breast tissue. According to The Merck Manual, more than 80% of breast cancer cases are discovered when the woman feels a lump [117]. Lumps found in lymph nodes located in the armpits can also indicate breast cancer. While 'manual' screening techniques are useful in determining the possibility of cancer, further testing is necessary to confirm whether a lump detected on screening is cancer, as opposed to a benign alternative such as a simple cyst. In a clinical setting, breast cancer is commonly diagnosed using a "triple test" of clinical breast examination (breast examination by a trained medical practitioner), mammography, and fine needle aspiration cytology. Both mammography and clinical breast exam, also used for screening, can indicate an approximate likelihood that a lump is cancer, and may also identify any other lesions. Fine Needle Aspiration and Cytology (FNAC), which may be done in a General Practitioner's office using local anaesthetic if required, involves attempting to extract a small portion of fluid from the lump. Clear fluid makes the lump highly unlikely to be cancerous, but bloody fluid may be sent off for inspection under a microscope for cancerous cells. Together, these three tools can be used to diagnose breast cancer with a good degree of accuracy [117].

Several treatments are available for breast cancer patients, depending on the stage of the cancer. Doctors usually take many different factors into account when deciding how to treat breast cancer. These kinds of factors may be the patient's age, the size of the tumour, the type of cancer a patient has, and many more. To get an idea of how a particular treatment may work for a specific patient and to predict how long the person may live, an indicator, called the Nottingham Prognostic Index (NPI) [64], has been introduced. It is not possible to predict exactly what will happen to each individual person, but this index can provide a general guidance. The NPI has been defined considering three factors

(defined below), which are combined using the formula:

$$\text{NPI} = (0.2 \times \text{tumour diameter in cms}) + \text{lymph node stage} + \text{tumour grade.}$$

The factors involved in the NPI definitions are [180]:

- The size of the cancer
- Whether the cancer has spread to the lymph nodes (lymph glands) under the arm (and if so, how many nodes are affected) – lymph node stage
- The grade of the cancer.

A more detailed description of the NPI and its clinical use may be found in Section 3.1.1, while the definition of grade is introduced in Section 7.3.

Cancer research produces huge quantities of data that serve as a basis for the development of improved diagnosis and therapies. Advanced statistical and machine learning methods are needed for interpretation of primary data and generation of new knowledge needed for the development of new diagnostic tools, drugs, and vaccines. Identification of functional groups and subgroups of genes responsible for the development and spread of this type of cancer as well as its subtypes are urgently needed for proper classification and identification of key processes that can be targeted therapeutically. In addition, accurate diagnostic techniques could enable various cancers to be detected in their early stages and, consequently, the appropriate treatments could be undertaken earlier [175].

Gene expression profiling using microarray has become a routine research tool in biomedicine. This high-throughput technology allows the researcher to monitor whole-genome gene expression profiles under different experiment conditions or disease phenotypes (including subtypes), as well as time course experiments. This type of screening helps to identify genes with similar expression pattern under various conditions or time course. Co-expression of genes often indicates their co-regulation or participation in related functional biological pathways or processes. One of the strategies in identifying

these groups of genes with similar expression patterns is by using unsupervised machine learning techniques such as clustering approaches.

Clustering is a multivariate analysis technique which aims to discover groups of similar objects within data [14]. An object is described either by a set of measurements or by relationships between the object and other objects. Cluster analysis does not use category labels that tag objects with prior identifiers. The absence of category labels distinguishes cluster analysis from discriminant analysis (and pattern recognition and decision analysis). The objective of cluster analysis is simply to find a convenient and valid organisation of the data, not to establish rules for separating future data into categories [88]. Clustering algorithms are used to find structure in the data.

Many clustering algorithms have been developed and applied to the analysis of microarray data ranging from simple hierarchical clustering to partitional approaches. The former seeks to build a hierarchy of clusters using either an agglomerative or divisive strategy. In the agglomerative approach, each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. In the divisive strategy, all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Hierarchical clustering algorithms are widely used, particularly in the medical field, as they produce a 'visual result': a tree (called dendrogram) depicting specified relationships among the entities [7] may be drawn to illustrate the arrangement of the clusters produced. On the other hand, partitional clustering produces clusters by optimising a criterion function defined either locally (on a subset of the data) or globally (defined over all of the patterns).

Partitional methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive [89]. All the clustering approaches mentioned so far are usually referred to as 'heuristic' approaches. However, a different kind of algorithms (model-based techniques) are also used to identify groups in data. Model-based clustering assumes that data are generated by a model and tries to recover the original model from the data themselves. The model recovered from the

data then defines clusters and an assignment of points to clusters [116]. Sometimes, model-based approaches assume that the model underlying data is in reality a mixture of probability distributions such as multivariate normal distributions [9]. All these kinds of analysis usually provide a good starting point for further examination of specific pathways and relevant biological processes. However, the complexity of this problem requires fine tuning and human intervention for determination of useful gene clusters associated with specific biological functions.

Other methods of classification (learning) are the so-called 'supervised algorithms'. Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown [101].

Algorithms for supervised learning range from decision trees to artificial neural network and from support vector machines to Bayesian classifiers. Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. Learned trees can also be re-represented as sets of if-then rules to improve human readability [128].

Artificial Neural Networks (ANNs) provide a general, practical method for learning real-value, discrete-valued, and vector-valued functions from examples. For certain types of problems, such as learning to interpret complex real-world sensor data, artificial neural networks are among the most effective learning methods currently known [80, 128]. However, especially for big data sets, ANNs may become huge and produce set of rules which are then difficult to interpret, especially for those researchers not familiar with computational analysis.

Support Vector Machines (SVMs) can also be used for pattern classification and non-linear regression. The main idea of a SVM is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximised in multi-dimensional space. The support vector machine can provide a good generalisation performance on pattern classification problems despite the fact that it does not incorporate problem domain knowledge [80].

Bayesian classifiers are based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data. In addition, Bayesian learning provides a quantitative approach to weighing the evidence supporting alternative hypotheses [128].

The whole range of classification techniques (supervised and unsupervised) may well help the identification of representative groups and subgroups of any kind of cancer and subsequently guide clinicians to the favourite and most powerful treatment. In particular, in such a multi-disciplinary work, the development of a new framework for elucidating core classes in a dataset may help in changing the approach clinicians usually use when dealing with breast cancer studies. This research project was in fact motivated by a previous investigation on breast cancer phenotypes using tissue microarray technology that was carried out by Mrs. Dalia M. Abd El-Rehim *et al.* from the Departments of Histopathology and Surgery, The Breast Unit, Nottingham City Hospital and University of Nottingham, UK, in 2005 [1]. Indeed, a single hierarchical clustering algorithm was used by the authors to categorise patients in different groups. Following this study, a co-operation with Professor Ian O. Ellis and his research group at the School of Molecular Medical Sciences at The University of Nottingham has been established to obtain and investigate a larger set of data relating to patients presented at Nottingham City Hospital between 1986 and 1998. A cohort of 1,076 women with primary operable invasive breast cancer has been made available for a multi-steps analysis in order to refine the phenotypic characterisation of the disease.

1.2 Aims and objectives of this Thesis

The ultimate goals of this multi-disciplinary research project concern both the clinical and technical aspects. From a clinical point of view, this work aims to move the field of medical decision making from the widely used approach of considering just a single classification technique (usually hierarchical clustering) to a multi-technique analysis one, in which different methods are investigated and results are derived from a consensus between techniques. Considering a more technical and computational aspect, the aim is to develop an original framework to elucidate core representative classes in a given dataset, in order to provide a step-by-step guideline which can be applicable across application domains (with a specific focus on breast cancer data sets).

Several research questions and hypotheses underlying the overall work were identified at the beginning of the project. Starting from the already published literature on breast cancer studies, it was noted that an extended review and comparison of different clustering algorithms had not been carried out yet. This gap of knowledge led to the formulation of the following research questions:

- Can a multi-techniques approach provide more accurate classification of breast cancer patients into sub-types of the disease?
- Is there a way to combine the results obtained by this multi-techniques approach?
- Is it possible to find an automated way to categorise a patient and give her immediately the most useful treatment?
- From a clinical point of view, can the medical decision making process be moved in the direction of personalised healthcare? If so, how should doctors be trained for that?

In order to achieve the aims stated above and to answer the research questions, the following objectives were identified:

(i) To establish standard methodologies for the acquisition, storage and extraction of electronic patient data.

(ii) To collect the various data sets available at the centres into a shared big resource, of unprecedented size, for breast cancer data.

(iii) To investigate different computational analysis methods (clustering) applicable across bioinformatics data and routine clinical information, including pathological and radiological data.

(iv) To determine an effective method to evaluate clustering results and to combine them in a set of representative groups of different cancer characteristics.

(v) To develop an automated supervised classification algorithm to be applied to any possible source of data, independent of their underlying distributions.

(vi) To combine the previously analysed techniques in an original framework to determine and emphasise core representative classes and validate this framework over novel data.

The first two objectives are discussed in Chapter 3, the third and fourth will be investigated in Chapter 4, and the fifth in Chapter 5. The last objective is discussed in Chapter 7. In the next section a more detailed structure of this work is presented.

1.3 Thesis Organisation

This thesis is structured as follows. Chapter 2 presents a literature review of various clustering approaches developed in the past to categorise data points in groups with high similarity. Both partitional and fuzzy methods are reported. A review of different clustering methods used in literature to group breast cancer data has been performed and is reported as well. Cluster validity is introduced in this chapter as a technique to assess the quality of clustering results and as a method to select the best number of groups to consider in the analysis. Several validity measures used in this thesis work are reviewed and

analysed in detail. In addition, a general description of techniques developed for consensus clustering is reported, explaining various methods to assess the comparison and the concordance among different clustering approaches. To conclude the overview on clustering algorithms, the model-based approach is described and the most commonly used algorithm, Expectation Maximisation, is presented. The chapter ends with a review on supervised classification methods, which are used to build models of the class distribution labels and to predict the class assignment of possible new objects.

In Chapter 3 a description of the medical background of this research work is presented. The disease of breast cancer is defined and its possible treatments are reviewed. Tissue samples of breast cancer were used for this study and in this chapter their collection and preparation, together with the instruments used for this purpose, are reported. The basic steps of data pre-processing are also given in Chapter 3, resulting in the development of a web interface for data storage and acquisition.

Chapter 4 is dedicated to the comparison of several clustering techniques applied on breast cancer data, namely hierarchical, fuzzy C-means, K-means, Partitioning Around Medoids (PAM) and Adaptive Resonance Theory (ART). Comparing different results using statistical analysis and visualisation techniques, an informal consensus clustering is defined resulting in the development of a set of six breast cancer core classes. Three of them present luminal characteristics (luminal biomarkers are over-expressed), differentiated by the presence or absence of other markers like estrogen receptors and/or progesterone receptors. In one of the six classes the HER2 marker (a human epidermal growth factor which gives higher aggressiveness in breast cancers) is strongly expressed. The last two classes are characterised by the over-expression of basal markers and the subsequent under-expression of the luminal ones. These two groups differ by the presence or absence of a marker called p53. These six classes are also characterised in terms of clinical outcome, highlighting some novel interesting results which have not previously been emphasised in literature. In addition, the work presented here supplies the evidence that cluster analysis should be treated with caution, as different algorithms produce different

groupings and it is always difficult to choose the best one. The chapter ends with a short note of interesting results on triple-negative cancers recently being found across several different breast cancer data sets analysed.

In Chapter 5 supervised classification methods are used to validate the classification obtained in Chapter 4 and to build a model for future possible patients entering the study. A comparison between three different techniques is performed and then a new 'non-parametric' algorithm is developed and presented to cope with the non-normality of most of the real world data sets. This new method is validated over known case studies taken from the Machine Learning Repository. Interesting results emerge, in particular in comparison with Multinomial Logistic Regression.

Chapter 6 provides a description of a recently proposed clustering algorithm called Affinity Propagation, which combines characteristics of both heuristic and model-based approaches and has the ability to suggest the number of clusters too. The main differences with other clustering algorithms like K-means are reported and then results from the application of Affinity Propagation over breast and cutaneous cancers are presented. In order to choose whether to include this algorithm in the proposed framework for core classes detection, a comparison of the CPU times needed for the computation of both Affinity Propagation and K-means was also performed. Several groupings known from literature were confirmed by this recently proposed method; in addition, this technique also suggested novel possible classifications for two of the data sets analysed. However, the time requested for computation was much longer than the time needed for the K-means one.

A step-by-step guide to identify core characteristic classes within any dataset is presented in Chapter 7. Starting with the application of different clustering algorithms and passing through several statistical methods and visualisation techniques to characterise the results, a set of core classes may be defined by a form of consensus clustering. It is then possible to verify these classes by using several supervised classification algorithms like decision trees, in order to obtain a set of rules which may be used for new data points in the future. This proposed framework is finally validated over a novel set of

histone markers for breast cancer patients. Results are still being analysed and verified by pathologists and researchers at the University of Nottingham, but, from a clinical point of view, the identified groups clearly distinguish patients with poor overall survival from those with low grading score and better survival rate. Considering the technical aspect, the identified classes result to be well separated and characterised by low, medium and high levels of biological markers.

The last chapter of this thesis, Chapter 8, concludes the work, drawing the main contributions and highlighting several possible directions for future research. A list of publications and oral presentations derived from this work thesis is reported at the end of the chapter.

This work was conducted as a joint research initiative between the University of Nottingham and the Institute of Medical Statistics, University of Milan and National Cancer Institute, Italy. The project itself was part of the BIOPTRAIN consortium, funded in the EU 6th Framework (FP6) as a Marie Curie Early Stage Training (EST) programme. The BIOPTRAIN mission was to establish a permanent European multi-centre interdisciplinary research training programme of world-class quality in bioinformatics optimisation algorithms.

1.4 Contributions to knowledge

Three main contributions to knowledge may be extracted from the work presented in the following chapters. Firstly, the novel framework to elucidate core representative classes in a dataset is an original methodological approach and its validation over a novel case study offers relevant results from both the technical and clinical perspectives.

Moreover, using different clustering techniques and a consensus between the resulting classifications, six diverse breast cancer groups are discovered and their characterisation appears to be novel in literature, especially in terms of their different clinical outcome.

Finally, a ‘non-parametric’ Bayesian classifier is developed to address the non norma-

lity of the data in many real world case studies. This novel algorithm is validated over different data sets from the Machine Learning Repository and results confirm its effectiveness when data is not normally distributed.

This research work led to several refereed papers – five journal papers (three accepted / published and two submitted) and four conference papers. A complete list of publications is reported in Section 8.4.

Chapter 2

Literature Review

This chapter provides a literature review of general clustering techniques as well as clustering approaches in breast cancer studies. As real medical applications are analysed in this thesis, clustering algorithms that can group breast cancer data are also introduced. One of the contributions of this work is the use of a consensus clustering approach for the identification of breast cancer classes which remain stable across different methods. For this reason, previously published literature on consensus clustering will be discussed. Later in this chapter, literature reviews on supervised classification and on model-based clustering will be presented.

The aim of this chapter is to present relevant background information about all the research subjects which have been used in the development of the original framework and to point out the gaps in the body of knowledge. This emphasises the motivation of the thesis: to develop a framework to elucidate core classes in a dataset which can be used for any available source of data.

2.1 Clustering techniques

Clustering is the process of grouping a set of unlabelled multidimensional patterns (objects or data points), such that patterns in the same cluster have the most similar characteristics, and patterns within different clusters have the most dissimilar characteristics.

In most cases a cluster is represented by a cluster centre or a centroid [89, 175]. Clustering has been applied to a wide range of applications, such as pattern recognition, image segmentation, spatial data analysis, machine learning, data mining, etc. Classification, another data analysis method, is often confused with clustering. The distinction between the two approaches is that classification is a supervised learning process which is trained on a set of pre-labelled patterns in order to predict into which class new patterns should be placed. In contrast, clustering is unsupervised, has no predefined classes and does not involve training examples [66, 90, 175]. As mentioned before, the aim of clustering is to group the patterns into clusters based on their similarity, which is defined by a distance measure.

A basic pipeline of a general clustering process can be described in the following way [89, 90]:

1. *Feature selection and/or extraction*. Feature selection is the process of identifying the most effective subset of the original features to use in clustering. Feature extraction is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering [89]. The purpose of this step is to make the clustering process work more efficiently as only the most important characteristics need to be considered. The objective is usually to reduce the time required for the clustering process without adversely affecting the quality of the clusters obtained [175].

2. *Pattern proximity measure appropriate to the data domain*. This measure is used to evaluate the similarity (or dissimilarity) of two data points. A variety of distance measures may be used depending on the problem under investigation [7, 89]. The most used ones are the Euclidean distance and correlation coefficients.

3. *Clustering or grouping*. The grouping step can be performed in a number of ways. The output clustering (or clusterings) can be hard (a partition of the data into

groups) or fuzzy (where each pattern has a variable degree of membership in each of the output clusters). Several of these approaches will be described in this chapter.

4. *Assessment of output* (if needed). In general, cluster validation may answer, among other aspects, questions such as: 'How good are the partitions?' 'Is there a better partitioning possible?', etc. The assessment of a clustering procedures output obtained in step 3 may be performed using cluster validity analysis. Several validity indices have been introduced in the past to assess the robustness and separation of clusters returned by clustering techniques [183].

In general, clustering techniques can be divided into two main categories, namely hierarchical and partitional clustering [7, 89]. In each category, many subtypes and variants have been applied to diverse types of clustering problems. In conventional clustering algorithms, each pattern has to be assigned exclusively to one cluster and the usual process is to optimise an objective function which somehow reflects the quality of the clusters. Where the physical boundaries of clusters are well defined, this approach can work well. However, when using data from real world applications, the boundaries between clusters might be vague. For this reason, fuzzy clustering extends the traditional clustering concept by allowing each pattern to be assigned to every cluster with an associated membership value. Therefore, in case of unclear cluster boundaries, fuzzy clustering may obtain more useful results.

In the following subsections, the key literature pertaining to both hierarchical and partitional clustering is identified and analysed.

2.1.1 Hierarchical

Hierarchical clustering is a way to group the data in a nested series of clusters [89]. The output of hierarchical clustering is a cluster tree, termed a *dendrogram*, which represents the similarity level between all of the patterns. Figure 2.1 shows a two-dimensional dataset which contains three clusters (data points have been labelled A – G). The den-

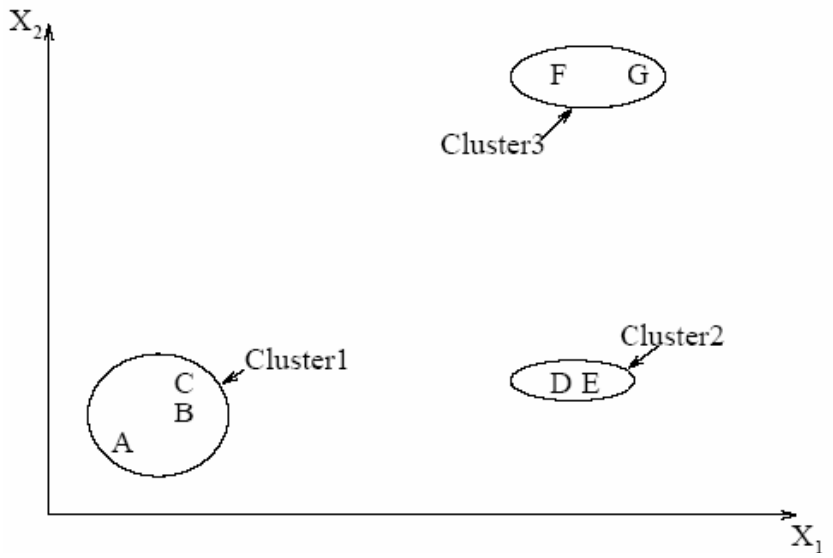


Figure 2.1: Two dimensional dataset with 3 clusters [89]

drogram corresponding to this figure has been displayed in Figure 2.2 [89]. As shown in Figure 2.2, a specific number of clusters can be generated through the vertical positioning of the cut-off line (dashed line in the figure). All of the data connected to the vertical line intersected by the cut-off line, belong to one cluster [175]. The position of the cutoff line is normally subjective and is decided based on the solution requirements. It should be noted that if the cut-off line is placed higher on the diagram, the total number of clusters is reduced, whereas, if the cut-off line is lowered, more clusters are produced.

To implement the standard method for the analysis of general data, one first constructs a dissimilarity measure for each pair of objects, often a distance measure. Alternatively, the dissimilarity measure may be taken to be one minus some measure of association, typically the correlation coefficient ρ [70]. Based on its algorithmic structure and operation, hierarchical clustering can be further categorised into agglomerative algorithms and divisive algorithms [89, 90]. The agglomerative method initially considers each of n patterns as an individual cluster and then the closest pair of distinct clusters is found and merged, leaving $(n-1)$ singleton clusters and one cluster with two distinct objects. The dissimilarity matrix is updated to take into account the merging that has occurred; based on the new dissimilarity matrix, the two closest distinct clusters are found and merged;

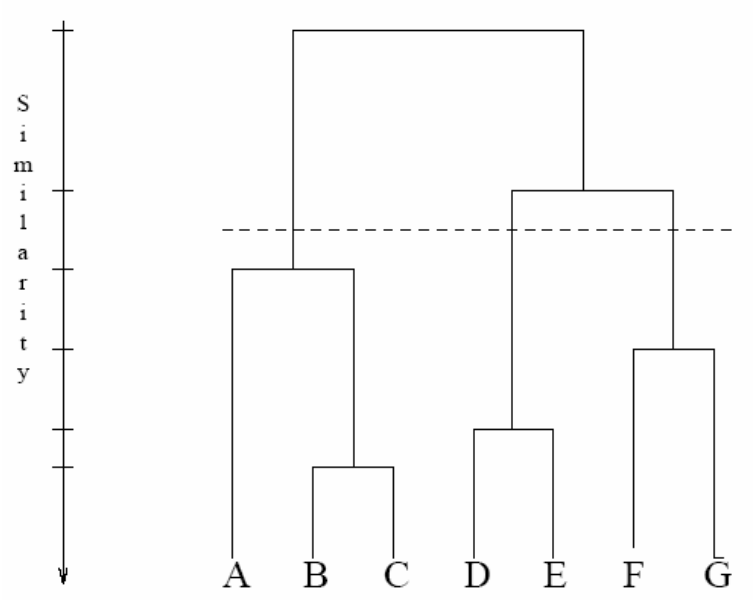


Figure 2.2: Dendrogram obtained from Figure 2.1 [89]

this continues until one cluster consisting of all n objects remains [70]. The dendrogram in the agglomerative approach is generated in a bottom-up fashion. In contrast, the divisive method starts by considering all n patterns in one cluster and, at each step, splits the cluster into two groups based on the similarity within the patterns, such that patterns in the same group have the highest similarity and patterns in the different groups have the most dissimilarity. In particular, in the first step, the object in the cluster that has the greatest dissimilarity to the other elements (the seed) is separated to form a so-called splinter group and the remaining elements in the original cluster are examined to see whether or not additional elements should be added to the splinter group. Two clusters result. The diameter of each cluster (the largest distance between observations in the same cluster) is then computed to see which is greater. The steps above are repeated with the cluster that has the greater diameter. This process is iterated until there are n singleton clusters [70]. The divisive approach is based on top-down dendrogram generation [175].

As part of the agglomerative algorithm, the linkage method provides a way to measure the similarity of clusters based on the patterns in the cluster [66]. The main linkage methods include single linkage, complete linkage, average linkage and minimum-variant

(Ward) algorithms [89, 90, 177]. Most of the other linkage methods are variants of these three. In the single linkage algorithm, the distance between two clusters is measured by the two closest patterns within the different clusters, thus resulting in the distance between two clusters being the minimum of all possible pairwise distances [70]. By contrast in the complete linkage algorithm, the distance between two clusters is measured by the two furthest patterns within the different clusters (the maximum between all possible pairwise distances). In average linkage clustering, the distance between two clusters is the average of the pairwise distances between two elements, one from the first cluster and the other from the second [70]. The minimum-variant algorithm is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimise the sum of squares of any two (hypothetical) clusters that can be formed at each step [6].

Depending on the linkage method employed, hierarchical clustering can generate clusters having different characteristics. For example, the single linkage algorithm has a tendency to produce a cluster with an elongated and irregular shape whereas the complete linkage algorithm can produce tight, compact and roughly hyper-spherical clusters [89], while the minimum-variant algorithm, although being regarded as very efficient, often tends to create clusters of small size [6].

A major drawback of hierarchical clustering methods is that group assignment of objects cannot change once an object has been placed in a cluster. These methods cannot undo what has been done in previous steps. In contrast, partition methods can reconsider cluster assignments at every stage [70].

2.1.2 Partitional clustering and K-means

In contrast to hierarchical clustering methods that produce a nested series of partitions, a method like K-means produces only a single partition [89]. Such, so called ‘partitional methods’, have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive. The partitional techniques

usually produce clusters by optimising a criterion function defined either locally (on a subset of the patterns) or globally (defined over all of the patterns). Combinatorial search of the set of possible labelings for an optimum value of a criterion is clearly computationally prohibitive. In practice, therefore, the algorithm is typically run multiple times with different starting states, and the best configuration obtained from all of the runs is used as the output clustering [89].

The most intuitive and frequently used criterion function in partitional clustering techniques is the squared error criterion, which tends to work well with isolated and compact clusters. Consider a data set $X = \{x_1, x_2, \ldots, x_n\}$, containing n patterns, which is to be clustered into c groups. Let us call $V = \{v_1, v_2, \ldots, v_c\}$ the corresponding set of centres and c_j the number of patterns in cluster j; let us also assume that each pattern can only belong to one cluster. The squared error e^2 can be expressed as

$$e^2 = \sum_{j=1}^{c} \sum_{i=1}^{c_j} ||x_i - v_j||^2 \tag{2.1}$$

where x_i is the i-th pattern in the j-th cluster, v_j is the j-th cluster centre, and $||x_i - v_j||$ is the Euclidean distance between x_i and v_j [89].

The K-means method is one of the early established algorithms in partitional clustering and is the simplest and most commonly used algorithm employing a squared error criterion [112]. The aim of the algorithm is to minimise the squared error criterion e^2 in Equation (2.1). It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centres until a convergence criterion is met (e.g., there is no reassignment of any pattern from one cluster to another, or the squared error ceases to decrease significantly after some number of iterations). The K-means clustering algorithm procedure may be described as follows:

(i) Choose c cluster centres to coincide with c randomly-chosen patterns or c randomly defined points inside the hypervolume containing the pattern set.

(ii) Assign each pattern to the closest cluster centre, according to the similarity measure

chosen.

(iii) Recompute the cluster centres v_j using the following formula:

$$v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i, \quad j = 1, \ldots, c \tag{2.2}$$

where, as for Equation (2.1), c_j is the number of patterns in cluster j, x_i is the i-th pattern in the j-th cluster, and c is the total number of clusters.

(iv) Reassign each pattern to the closest cluster centre.

(v) If a convergence criterion is not met, go to step (iii). Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centres, or minimal decrease in squared error.

2.1.3 Partitioning Around Medoids (PAM)

When partitioning a set of objects into k clusters, the main objective is to find clusters of objects which show a high degree of similarity, while objects belonging to different clusters are as dissimilar as possible. Of course, many methods exist that try to achieve this aim [97]. The PAM algorithm (Partitioning Around Medoids) is based on the search for *k representative objects* among the objects of the data set. As evoked by their name, these objects should represent various aspects of the structure of the data. In the cluster analysis literature such representative objects are often called *centrotypes*. In the PAM algorithm the representative objects are the so-called *medoids* of the clusters and the aim of PAM is to minimise the average dissimilarity of objects to their closest medoids [97]. Dissimilarities are nonnegative numbers $d(i, j)$ that are small (close to zero) when two data points i and j are 'near' to each other and become large when i and j are very different. It is usually assumed that dissimilarities are symmetric and that dissimilarity of an object to itself is zero, but in general the triangle inequality does not hold [97]. In general, a Euclidean metric is used for calculating dissimilarities between observations.

After finding a set of k representative objects, the k clusters are constructed by assigning each object of the data to the nearest representative object.

In many clustering problems one is particularly interested in a characterisation of the clusters by means of typical or representative objects. These are objects that represent the various structural aspects of the set of objects being investigated. There can be many reasons for searching for representative objects. Not only can these objects provide a characterisation of the clusters, but they can often be used for further work or research, especially when it is more economical or convenient to use a small set of k objects instead of the large set one started off with. In the method used in PAM, the representative object of a cluster is its medoid, which it is defined as that object of the cluster for which the average dissimilarity to all the objects of the cluster is minimal [97].

In addition, the results of partitional techniques like K-means and PAM are lists of clusters with their objects, which are not as visually appealing as the dendrograms of hierarchical methods [150]. In order to obtain a graphical representation of each clustering for such methods, Rousseeuw suggested to plot the so-called silhouettes, which were introduced by himself in [150]. Moreover, the silhouettes plot can be used to select the number of clusters and to assess how well individual observations are clustered [42]. Let a_i denote the average dissimilarity between i and all other observations in the cluster to which i belongs. For any other cluster C, let $d(i,C)$ denote the average dissimilarity of i to all object of C and let b_i denote the smallest of these $d(i,C)$. The silhouette width of observation i is

$$s_i = (b_i - a_i)/\max\{a_i, b_i\}$$

and the overall average silhouette width is simply the average of s_i over all the observations i,

$$\bar{s} = \frac{1}{n}\sum_i s_i.$$

Intuitively, object with large silhouette width s_i are well clustered, whereas those with small s_i tend to lie between clusters [42]. Kaufman and Rousseeuw suggest estimating

the number of clusters k by that which gives the largest average silhouette width $\bar{s}$ [97]. From the preceding definition, it is clear that

$$-1 \leq s_i \leq 1$$

for each observation i. When s_i is at its largest (close to 1) it can be said that i is 'well classified', as there appears to be little doubt that i has been assigned to an appropriate cluster. A different situation occurs when s_i is about zero. Then a_i and b_i are approximately equal and hence it is not clear whether i should have been assigned to which cluster. The worst situation takes place when s_i is close to -1. Then a_i is much larger than b_i, so i lies on the average much closer to another cluster rather than the one it has been assigned to. Therefore it would have seemed more natural to assign object i to another cluster, so it can be almost concluded that observation i has been 'misclassified' [97].

Having computed the quantities s_i from the dissimilarities, the graphical display can be constructed. The silhouette of a generic cluster C is a plot of the s_i, ranked in decreasing order, for all observations i in C. For each observation i, a bar is drawn, representing its silhouette width s_i. Observations are grouped per cluster, starting with cluster 1 at the top. Observations with a large s_i (almost 1) are very well clustered, a small s_i (around 0) means that the observation lies between two clusters, and observations with a negative s_i are probably placed in the wrong cluster. An example of the silhouette plot is shown in Figure 2.3.

2.1.4 Fuzzy C-means

Fuzzy sets are an extension of classical set theory and are used in fuzzy logic. In classical set theory the membership of elements in relation to a set is assessed in binary terms according to a crisp condition – an element either belongs or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in relation to a set; this is described with the aid of a membership function

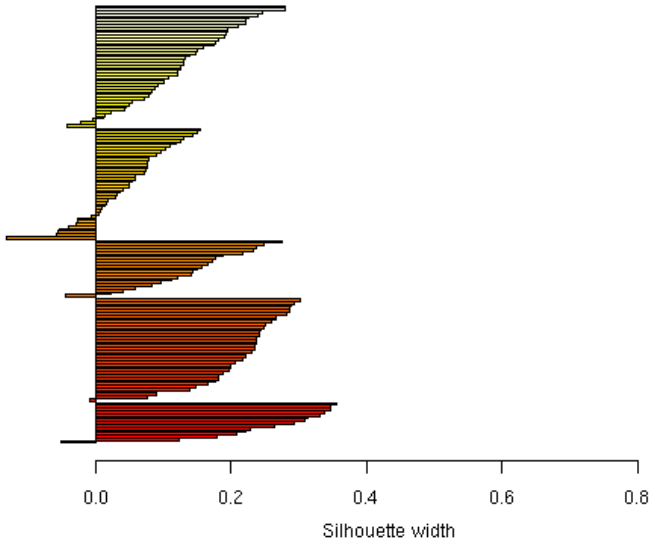


Figure 2.3: Silhouette plot for a clustering technique

$\mu \rightarrow [0,1]$. Fuzzy sets are an extension of classical set theory since, for a certain universe, a membership function may act as an indicator function, mapping all elements to either 1 or 0, as in the classical notion.

Gitman and Levine [69] appear to have been the first to apply the theory of fuzzy sets to clustering problems. They developed an algorithm which partitions a given sample from a "multimodal fuzzy set" into "unimodal fuzzy sets" [69]. The notion of a unimodal fuzzy set has been chosen by the authors to represent the partition of a data set for two reasons. First, it is capable of detecting all the locations in the vector space where there exist highly concentrated clusters of points, since these will appear as modes according to some measure of 'cohesiveness'. Second, the notion is general enough to represent clusters which exhibit quite general distributions of points [69]. An important feature of this method was the presence of a further 'dimension', the order of 'importance' of every point, as an aid in the clustering process. This is accomplished by associating with every point in the set a grade of membership or characteristic value [195]. Thus the order of the points according to their grade of membership, as well as their order according to distance, are used in the algorithm. The latter partitions a sample from a multimodal fuzzy set into unimodal fuzzy sets [69].

One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) algorithm, which was firstly developed by Dunn in 1974 [44] and subsequently improved by Bezdek in 1981 [12]. Bezdek's improvement consisted in the introduction of the so-called *fuzzifier* parameter m. The algorithm is based on the minimisation of the fuzzy objective function

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{i,j})^m \|x_i - v_j\|^2. \tag{2.3}$$

Once again, as in Section 2.1.2, $X = \{x_1, x_2, \ldots, x_n\}$ represents a collection of n data points and $V = \{v_1, v_2, \ldots, v_c\}$ is the corresponding set of cluster centres. In addition, $\mu_{i,j}$ is the membership degree of data x_i to the cluster centre v_j and it must satisfy the two following conditions:

$$\mu_{ij} \in [0,1],\ i = 1,\ldots,n \text{ and } j = 1,\ldots,c, \tag{2.4}$$

$$\sum_{j=1}^{c}\mu_{ij} = 1. \tag{2.5}$$

Parameter m is called the 'fuzzifier' or 'fuzziness index' and is used to control the fuzziness of the membership of each data point. When $m = 1$ the fuzzy c-means is equivalent to the K-means algorithm and the larger the value of m, the fuzzier the method becomes. Even though there are no theoretical basis for the optimal selection of the fuzziness index m, a value of $m = 2.0$ is usually chosen [12]. All the membership degree values from each observation to all cluster centres form the fuzzy partition matrix $U = (\mu_{ij})_{n\times c}$. The fuzzy c-means algorithm procedure may be described as follows [175]:

(i) Fix the number of cluster c $(2 \leq c < n)$ and initialise the fuzzy partition matrix U with a random value such that it satisfies conditions (2.4) and (2.5).

(ii) Calculate the fuzzy centres v_j using the following formula:

$$v_j = \frac{\sum_{i=1}^{n}(\mu_{ij})^m x_i}{\sum_{i=1}^{n}(\mu_{ij})^m},\ \forall j = 1,\ldots,c. \tag{2.6}$$

(iii) Update the fuzzy partition matrix U using the following formula:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{\|x_i - v_j\|}{\|x_i - v_k\|})^{\frac{2}{m-1}}}, \; i = 1, \ldots, n \text{ and } j = 1, \ldots, c. \tag{2.7}$$

(iv) If a convergence criterion is not met, go to step (ii). Typical convergence criteria are that the difference between updated and previous objective function J is less than a prespecified minimum threshold or that the maximum number of iteration cycles has been reached.

The fuzzy c-means algorithm, as well as K-means, needs the number of clusters to be pre-specified in advance as an input parameter to the algorithm. Moreover, both these techniques begin with random initialisation of the cluster centres. This can lead the approaches to suffer premature convergence to local optima. If the initial cluster centres are not appropriate, the iterative improvement of the centre positions can result in locally optimal solutions being obtained [175]. Also for this reason in this work, as it will be described later on, the initial cluster centres for K-means and fuzzy c-means were set using a hierarchical approach.

Table 2.1 summarises the main concepts and characteristics of the clustering algorithms considered in this thesis. More specifically, the following features of each algorithm are reported: i) the algorithm complexity, ii) the input parameters, iii) the clustering criterion and iv) the results produced by the algorithm [74].

2.2 Clustering validation

In cluster analysis, one of the most important issues is the choice of the optimal number of cluster to consider. It is also important to assess the resulting clusters produced by a single algorithm.

Clustering validation is a concept that is used to evaluate the quality of clustering results. In addition, if the number of clusters within the data is not known prior to com-

Category	Partitional			
Name	Time Complexity[a]	Input parameters	Results	Clustering criterion
K-Means	$O(nkt)$	Number of clusters	Centres of clusters	$\min_{v_1,v_2,\ldots,v_k}(E_k)$ $E_k = \sum_{j=1}^{c}\sum_{i=1}^{c_j}\|\|x_{ij}-v_j\|\|^2$
PAM	$O(tk(n-k)^2)$	Number of clusters	Medoids of clusters	$\min(TC_{ih})$ $TC_{ih} = \sum_j C_{jih}$ C_{jih} = cost of replacing centre i with h as far as j is concerned
FCM	$O(nkt)$	Number of clusters	Centre of clusters	$\min_{U,v_1,v_2,\ldots,v_k}(J(U,V))$ $J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{i,j})^m\|\|x_i-v_j\|\|^2$
Category	Hierarchical			
Name	Time Complexity[a]	Input parameters	Results	Clustering criterion
HCA	$O(n^2)$	A set of dissimilarities for the objects being clustered	Cluster tree, the dendrogram	Minimise the sum of squares of any two clusters that can be formed at each step

[a] n is the size of the dataset, k is the number of cluster defined and t is the number of iterations.

Table 2.1: The main characteristics of the clustering algorithms analysed

mencing an algorithm, a cluster validity index may be used to determine the best number of clusters for the given data set. In general, cluster validation may answer, among other aspects, questions such as [148]:

1. How many cluster should be considered?
2. How good are the partitions?
3. Is there a better partition possible?

In general, cluster validity measures can be expressed in terms of the three types of criteria named below [88, 175].

External criteria measure performance by matching a clustering structure to a priori information. For example, an external criterion measures the degree of correspondence between cluster numbers, obtained from a clustering algorithm, and category labels, assigned a priori. An external criterion can also measure the degree to which data confirm a priori ideas without a formal cluster analysis being performed.

Internal criteria assess the fit between the structure and the data, using only the data themselves. For example, an internal criterion would measure the degree to which a partition, obtained from a clustering algorithm, is justified by the given proximity matrix.

Relative criteria decide which of two structures is better in some sense, such as being more stable or more appropriate for the data. For example, a relative criterion would measure quantitatively whether a single-link or a complete-link hierarchy fits the data better.

The fundamental idea of the first two types of approach (external and internal criteria) is to test whether the data points in the given dataset are randomly structured or not, based on statistical testing. This usually requires some sort of calculation involving pairwise comparison between each pair of data points and each cluster, which leads to a

computationally expensive procedure. In addition, the indices related to these approaches aspire to measure the degree of the dataset to a pre-specified clustering scheme [74, 175]. Conversely, the third approach does not involve statistical tests and allows for the best clustering structure to be chosen from a set of schemes, defined based on pre-specified criteria [74, 175].

2.2.1 Validity indices

In this thesis, real world problems from the medical domain are considered. In such problems the different types of phenotype (corresponding to the number of clusters from a clustering analysis point of view) are often unknown in advance. However, using a validity index, or a sort of consensus of several validity indices, the best clustering scheme may be identified. This can be implemented by applying the clustering techniques within a range of cluster numbers, and considering the partition with the best cluster validity index value. The whole procedure may be summarised by the following steps [110, 148, 175]:

1. For a given data source X, fix all the clustering parameters but the number of clusters c.

2. Set the values of the minimum and maximum number of clusters, respectively c_{min} and c_{max}.

3. For c running between c_{min} and c_{max} compute the following three steps:

   (a) Initialise the cluster centres.

   (b) Apply the clustering method with number of clusters c.

   (c) Calculate and store the validity index of the clustering scheme

4. Choose the clustering structure that corresponds to the best validity index value obtained throughout the procedure.

Although there are many variations of validity indices, they are all either based on considering the data dispersion in a cluster and between clusters, or considering the scat-

ter matrix of the data points and the one of the clusters centres. The remaining part of Section 2.2 is dedicated to the different types of validity indices, suitable for both the hard clustering and the fuzzy one.

2.2.2 Validity indices for hard clustering

Several validity measures were proposed in recent years. In a technical report Weingessel *et al.* [183] conducted an examination of 14 indices for determining the number of clusters on artificial binary data sets being generated according to various design factors and to resemble real-world data. To provide a variety of clustering solutions the data sets were analysed by different non hierarchical clustering methods. The purpose of the paper was to present the performance and the ability of an index to detect the proper number of clusters in a binary data set under various conditions and different difficulty levels. The indices reported in this section were all analysed in [183] and are those used in this thesis work.

TraceW Validity Index

In 1965, Edwards and Cavalli-Sforza [46] found that the analysis of variance provided an excellent criterion for testing the goodness of a particular cluster division. Using the analysis of variance technique, data points may be divided into the two most-compact clusters (clusters as dense as possible), and the process repeated sequentially so that a 'tree' diagram may be formed. When points are divided into two clusters the sum of the squared distances from their mean could be partitioned into the sum of the squared distances of the points of one cluster from *their* mean, the similar sum for the other cluster, and the between-clusters sum of squares [46]. Thus, Edwards and Cavalli-Sforza suggested that the natural criterion for division was clearly the between-cluster sum of squares and, to obtain the best split, this sum should be maximised (and the within-clusters sum of squares consequently minimised). The last criterion could also be presented as the minimisation of the 'Trace of W', where W is the pooled-within groups scatter ma-

trix [61], and it is defined in the following way. Suppose that data are given in the form of a matrix $(X)_{n\times p}$ with the ith row given by the $(l\times p)$ vector $P_i=(x_{i1},\ldots,x_{ip})$ representing the observation vector of the ith object. Suppose also to have a partition of n objects into g groups with $n_1,n_2,\ldots,n_g$ objects in each group and $n=\sum_{i=1}^{g}n_i$. Then for the kth group the row vectors P_{lk} for $l=1,\ldots,n_k$ represent the objects in group G_k. The scatter matrix of each group G_k with centre of gravity vector C_k is

$$W_k=\sum_{l=1}^{n_k}(P_{lk}-C_k)^T(P_{lk}-C_k).$$

The sum over all k

$$W=\sum_{k=1}^{g}W_k$$

defines what is called the pooled-within groups scatter matrix [61].

Friedman and Rubin Validity Indices

Friedman and Rubin [61] criticized the method proposed by Edwards and Cavalli-Sforza arguing that although Trace W was invariant under an orthogonal transformation, it was not invariant under any non-singular linear transformation. In addition, Friedman and Rubin defined the between groups scatter matrix by

$$B=\sum_{k=1}^{g}n_kC_k^TC_k,$$

where notation is the same as above. For each partition of the n objects into g groups the following matrix identity can be defined

$$T=W+B \tag{2.8}$$

where T is the total scatter matrix of the n points and it is given by

$$T=X^TX=\sum_{i=1}^{n}P_i^TP_i.$$

Using Equation (2.8) Friedman and Rubin stated that, since T is constant over all the partitions, minimising Trace W is equivalent to maximising Trace B, because Trace $T =$ Trace W + Trace B [61].

Then, two possible scenarios were considered, $p = 1$ (one variable) and $p > 1$ respectively. For $p = 1$, Equation (2.8) is a statement about scalars and since the total scatter T is fixed, a natural criterion for grouping is to minimise W. This is equivalent to maximising B. Also for $p = 1$, the following may be written $T/W = 1 + B/W$ where B/W multiplied by the ratio of the degrees of freedom is what in statistics is called an F ratio. This ratio is invariant under non-singular linear transformations of the data. The criterion may thus be restated as partitioning the n objects into g groups so as to maximise the ratio B/W or equivalently T/W.

For $p > 1$, Equation (2.8) is a matrix equation and the question of criteria for grouping, which are invariant under non-singular linear transformations of the original data matrix, is more complex. These criteria are derived from the identity $T = W + B$. One criterion is to maximise the ratio of determinants

$$\frac{|T|}{|W|} = |I + W^{-1}B|. \tag{2.9}$$

That is in principle all partitions of the n objects into g groups are considered and that partition into g groups for which this ratio is maximum is to be chosen. As $|T|$ is fixed, it is sufficient to minimise $|W|$ [61].

Another criterion function related to the basic identity (2.8) proposed by Friedman and Rubin was the maximum of the Trace$[W^{-1}B]$ over all partitions into g groups. The function Trace$[W^{-1}B]$ has been used as a test statistic in the same way as the ratio of the two determinants mentioned previously. Moreover, both Trace $W^{-1}B$ and $|T|/|W|$ may be expressed in terms of the eigenvalues of $W^{-1}B$. In particular

$$\frac{|T|}{|W|} = \prod_{i=1}^{t}(1 + \lambda_i) \text{ and,}$$

$$\text{Trace } W^{-1}B = \sum_{i=1}^{t} \lambda_i.$$

These eigenvalues are solutions of the determinant equation $|B - \lambda W| = 0$. All the eigenvalues of this equation are known to be invariant under non-singular linear transformations of the original data matrix. In fact they are the only invariants of W and B under such transformations [61].

In their work Friedman and Rubin assessed that if a single criterion function with which to explore the structure of heterogeneous multivariate data had to be chosen, they would choose $|T|/|W|$ since it is invariant under non-singular linear transformations and has demonstrated on the data analysed in their work a greater sensitivity to the local structure of data than the other criteria. In particular, groups resulting from the trace $W^{-1}B$ criterion were always separable by a single discriminant function (a single direction in space). This was not true for the $|T|/|W|$ criterion. Moreover, from a computational point of view, it is faster to compute $|W|$ than W^{-1}. As a matter of fact, the minimum Trace W criterion, proposed by Edwards and Cavalli-Sforza [46], is much less costly in computer time than any other criteria. However, its major fault is that it does not take into account the within-group covariance of the measurements.

Scott and Symons Validity Index

According to Scott and Symons [154], clustering techniques seem to be applied in two rather different situations. In one case, the purpose of the analysis is purely descriptive. There are no assumptions, implicit or otherwise, about the form of the underlying population and the grouping is simply a useful condensation of the data. In the other case, it is felt that the population is composed of several distinct sub-categories and the purpose of the analysis is to group together all those observations belonging to the same sub-category. Scott and Symons preferred to consider the second type of problems [154].

As a model for this situation, the authors supposed that each observation in the sample may arise from any one of a small number of different distributions. This would be the

standard classification problem if the distributions were known, or there was a substantial amount of information about them from previous samples, but little or no prior knowledge about the component distributions is available in most situations where clustering techniques are used. In either case, classification or clustering, the goal is to group together all the observations from the same distribution. Let denote γ the set of identifying labels, i.e., if there are n sample observations, γ is an unknown parameter with n components, where the ith component indicates the distribution from which the ith observation came. The maximum likelihood (ML) estimate of γ, under the assumption that the underlying distributions are multivariate normal, is derived and this turns out to be equivalent to several standard clustering methods with different assumptions about the covariance structure. These methods are shown to be natural extensions of standard classification rules based on the likelihood ratio criterion [154].

Supposing that the observations are drawn independently form a mixture of multivariate normal distributions leads to considering the model above with the additional assumption that γ is an (unobservable) random variable whose components are the outcomes of n independent multinomial trials. An indirect estimate of γ is obtained by estimating the parameters of the mixture and using standard classification methods with these estimates in place of the unknown parameter.

The model proposed by Scott and Symons is now described. The sample consists of n observations $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, where $\mathbf{y}_i$ represents measurements on p characteristics. Suppose that the observations are independent and that each may arise from any one of G possible p-variate normal distributions with means $\mu_1, \ldots, \mu_G$ and covariance matrices $\sigma_1, \ldots, \sigma_G$. To be as general as possible, it is allowed for the possibility of a previous sample of independent observations $\mathbf{x}_{g_1}, \ldots, \mathbf{x}_{g_{m_g}}$ from each distribution. Then the joint distribution of $\mathbf{Y}$ and the previous observations is completely determined by μ_g, σ_g ($g = 1, \ldots, G$) and the grouping or classification parameter $\gamma = (\gamma_1, \ldots, \gamma_n)$, where $\gamma_i = g$ when y_i comes from the g-th sub-population. If $\theta = (\gamma, \mu_1, \ldots, \mu_G, \sigma_1, \ldots, \sigma_G)$ denotes the

collection of all the parameters, the loglikelihood function, $l(\theta)$, is given by

$$l(\theta) = -\frac{1}{2}\sum_{g=1}^{G}\Big[\sum_{i=1}^{m_g}(\mathbf{x}_{g_i}-\mu_g)'\sigma_g^{-1}(\mathbf{x}_{g_i}-\mu_g) + \sum_{C_g}(\mathbf{y}_i-\mu_g)'\sigma_g^{-1}(\mathbf{y}_i-\mu_g) + (m_g+n_g)\log|\sigma_g|\Big], \quad (2.10)$$

where C_g is the set of $\mathbf{y}_i$'s assigned to the gth group or cluster by γ, n_g is the number of observations in C_g, and $|\sigma_g|$ denotes the determinant of σ_g [154].

The classification or clustering problem is to estimate γ and hence the clusters C_1, ..., C_G. If the means and covariances are known, or there are a large number of previous observations from each sub-population, this is the classical model for the classification problem. When there is little or no prior information about the components, the problem becomes one of cluster analysis [154].

The likelihood in Equation (2.10) can be maximised by substituting the ordinary ML estimates of μ_g and σ_g. The estimate of μ_g, whatever the assumption about the σ_g, is

$$\hat{\mu}_g(\gamma) = (m_g\bar{\mathbf{x}}_g + n_g\bar{\mathbf{y}}_g)/(m_g+n_g),$$

where $\bar{\mathbf{y}}_g$ is the mean of the n_g observations in C_g. When $\hat{\mu}_g(\gamma)$ is substituted for μ_g in expression (2.10) it follows that the ML estimate, $\hat{\gamma}$, of γ can be found by minimising

$$\sum_{g=1}^{G}\{\mathrm{tr}[(\mathbf{W}_{g_x}+\mathbf{W}_{g_y}+\mathbf{W}_{g_{xy}})\sigma_g^{-1}] + (m_g+n_g)\log|\sigma_g|\} \quad (2.11)$$

where

$$\mathbf{W}_{g_x} = \sum_{i=1}^{m_g}(\mathbf{x}_{g_i}-\bar{\mathbf{x}}_g)(\mathbf{x}_{g_i}-\bar{\mathbf{x}}_g)', \qquad \mathbf{W}_{g_y} = \sum_{C_g}(\mathbf{y}_i-\bar{\mathbf{y}}_g)(\mathbf{y}_i-\bar{\mathbf{y}}_g)',$$

and

$$\mathbf{W}_{g_{xy}} = \frac{m_g n_g}{m_g+n_g}(\bar{\mathbf{y}}_g-\bar{\mathbf{x}}_g)(\bar{\mathbf{y}}_g-\bar{\mathbf{x}}_g)'.$$

Two different situations may occur: the covariance matrices are equal or are different.

In the first scenario, expression (2.11) reduces to

$$\text{tr}[(\mathbf{W}_x + \mathbf{W}_y + \mathbf{W}_{xy})\sigma^{-1}] + (m+n)\log|\sigma|, \tag{2.12}$$

where $\mathbf{W}_x = \sum \mathbf{W}_{g_x}$ is the within-groups sum of squares matrix for the $\mathbf{x}$'s, $\mathbf{W}_y = \sum \mathbf{W}_{g_y}$ is the within-groups sum of squares matrix for the $\mathbf{y}$'s, and $\mathbf{W}_{xy} = \sum \mathbf{W}_{g_{xy}}$ is the contribution due to the differences between the $\mathbf{y}$'s and $\mathbf{x}$'s. If σ is known, (2.12) reduces further and the criterion proposed by Edwards and Cavalli-Sforza [46] may be obtained as a particular example. If σ is not known, it follows that $\hat{\gamma}$ is the grouping that minimises

$$|\mathbf{W}_x + \mathbf{W}_y + \mathbf{W}_{xy}|.$$

A particular case of this scenario is the criterion proposed by Friedman and Rubin in [61].

When, instead, the covariance matrices are all different and are specified, Scott and Symons attributed little of interest about this case. On the other hand, if σ's are not known, $\hat{\gamma}$ is the grouping that minimises

$$\prod_{g=1}^{G} |\mathbf{W}_x + \mathbf{W}_y + \mathbf{W}_{xy}|^{m_g+n_g}.$$

It is important to realize that the maximum likelihood methods will always partition the data into the maximum number of partitions allowed without suggesting the best number of groups. Scott and Symons suggested to rephrase the question of 'how many clusters are there?' as a testing problem. For example, the fundamental question whether there is more than one cluster can be considered as a test of the null hypothesis $H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_n$ against the alternative that not all the γ_i's are equal. If λ denotes the likelihood ratio statistic, then, in the case $\sigma_g = \sigma(g = 1, \ldots, G)$ with σ unknown,

$$-2\log\lambda = n\log[\max_{\gamma}(|\mathbf{T}|/|\mathbf{W}_y|)], \tag{2.13}$$

where $\mathbf{T} = \sum(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ is the total scatter matrix [154].

The maximisation of the quantity on the right hand side of (2.13) is today known as the Scott and Symons criterion. The maximum likelihood approach and the assumption that data are generated by a mixture of underlying probability distributions will be used again in model-based clustering (see Section 2.5) where the criterion of Scott and Symons will be derived as a particular case.

Marriott Validity Index

Starting from the criterion proposed by Friedman and Rubin [61] for clustering validity, Marriott analysed it from the point of view of a practical user, stating that certain difficulties arise both in the computational and in the interpretation of the results [118]. According to his work, the basic question of 'how many groups are there for a set of observation?' may be answered in two stages:

(a) What is the best subdivision into a given number, g, of groups?

(b) What is the best value of g?

One possible answer to (a) is that subdivision which minimises the variability within groups. In terms of the procedure proposed by Friedman and Rubin [61] this means minimising the determinant of the variance-covariance matrix $|\mathbf{W}|$. However if one of the variates is strongly grouped, the optimum subdivision defined by this approach may well be entirely based on that variate. The method searches for *any* natural grouping, not necessarily one based on all the measurements. According to Marriott the main advantages of using $|\mathbf{W}|$ are that variables that are highly correlated in the whole population are not given excessive weight and that a grouping that depends on high correlations within groups is readily detected [118].

When an answer to the first point reported above is given for a range of values of g, from 1 up to some preassigned maximum, it is necessary to decide which, if any, of the subdivisions corresponds to a 'natural' clustering of the data. It is clear that if the data are drawn from a population which is strongly grouped round a small number of

modes, the optimum subdivision into the same number of groups will produce a large reduction in $|\mathbf{W}|$. If the underlying population is unimodal, or uniform, the reduction is likely to be much smaller. It is easy to see that the optimum subdivision into g groups of a uniformly distributed population reduces $|\mathbf{W}|$ by a factor g^2. This suggests that the criterion $g^2|\mathbf{W}|$ may provide an answer to point (b) above [118] and the minimisation of $g^2|\mathbf{W}|$ is nowadays known as the Marriott validity index. The maximum difference between successive levels is used to determine the best partition level [126].

As Marriott recognises in his original work, it is important to note that, although the criterion $g^2|\mathbf{W}|$ divides populations in a way that conforms reasonably well with intuitive ideas of natural clustering, if the population consists of a mixture of subpopulations with very different dispersion matrices, it may be split into too many subdivisions and recombination of some groups may be necessary [118].

From a theory prospective, if the observations are considered independently uniformly distributed on (0, 1), the subdivision of a univariate rectangular distribution on (0, 1) into g groups is optimum when the sections are equal. This reduces the standard deviation by a factor g, and the variance by g^2. Such a subdivision of any of the k covariates in the multivariate case reduces $|\mathbf{W}|$ by a factor of g^2. If g is prime, these are the only optimum subdivisions. If g is composite, however, there are further cases; if $g = g_1 g_2$, a division of any two covariates into g_1 and g_2 equal subdivisions respectively will also reduce $|\mathbf{W}|$ by g^2. In fact, all possible subdivisions into 'boxes' of the same size, shape, and orientation are optimum. It seems reasonable to adopt as an optimum subdivision of a distribution or sample the one that minimises $g^2|\mathbf{W}|$ [118].

Marriott also described those drawbacks of his criterion, especially those problems that may arise from the data. In particular, considering the computational procedure, he stated that, ideally, the optimum subdivision of a data set could be found by calculating the criterion $|\mathbf{W}|$ for all possible subdivisions into g groups, and selecting the least value [118]. In practice, however, this is only possible for trivially small data sets, except in the one-dimensional case.

Another issue concerns the way in which data are collected and recorded. If this procedure is done with limited accuracy it may happen that the grouping will be affected. In general, the less accurately a measurement is recorded, the more influence it is likely to have on the grouping [118].

The most significant problem is related to a possible linear dependence between covariates. In particular, when many variables are measured, often there are high correlations between them; in this case, the grouping procedure breaks down, and it is necessary to reduce the dimensionality of the data. The simplest way of doing so suggested by Marriott is by taking principal components. A preliminary analysis transforms the data into principal components, and then the first k are chosen for use in the cluster analysis. The value of k should be large enough to include virtually all the information in the original data; the process should stop when the variance of the components becomes comparable with the measurements error [118]. It is important to note that the actual vectors representing the principal components are not important in the subsequent analysis. What is important is the space they span, and any linear combination of them would give the same optimum subdivisions. Whereas the cluster analysis is scale-independent, the principal components are not. Provided k is sufficiently large, however, this is not important. Different methods of standardisation may give quite different component vectors, but the space that they span will be virtually the same. It is also worth noting that a small value of $|\mathbf{W}|$ does not necessarily imply that the component groups are unimodal. High correlation within groups may 'hide' a secondary grouping, and this can only be detected by further reduction in dimensionality.

Calinski and Harabasz Validity Index

Milligan and Cooper [126] have provided a survey of several validity indices for data sets containing distinct non–overlapping clusters while using only hierarchical clustering algorithms. The authors recommended to use the Calinski and Harabasz procedure which, for their experiment, was the best performing validity criterion.

This index [20] for n data points and K clusters is computed as

$$\frac{[traceB/(K-1)]}{[traceW/(n-K)]}.$$

Here, B and W are the between and within cluster scatter matrices. The maximum hierarchy level is used to indicate the correct number of partitions in the data. The trace of the between cluster scatter matrix B can be written as

$$traceB = \sum_{k=1}^{K} n_k \|z_k - z\|^2,$$

where n_k is the number of points in cluster k and z is the centroid of the entire data set. The trace of the within cluster scatter matrix W can be written as

$$traceW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_i - z_k\|^2$$

where z_k is the centroid of the cluster k [120]. Therefore, the Calinski and Harabasz (CH) index can be written as [120]

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \|z_k - z\|^2}{K-1}\right] / \left[\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_i - z_k\|^2}{n-K}\right].$$

According to Käster *et al.* [96], matrices B and W can be computed in the following way:

$$B = \sum_{C_i \in G} |C_i| (\mathbf{m}_i - \bar{\mathbf{x}})(\mathbf{m}_i - \bar{\mathbf{x}})'$$

and

$$W = \sum_{C_i \in G} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)',$$

where C_i indicates a cluster of the grouping G, $\mathbf{m}_i$ is the centroid of cluster C_i, and $\bar{\mathbf{x}}$ is the mean of all data points in the data set under investigation.

2.2.3 Validity indices for fuzzy clustering

In fuzzy clustering methodology, the fuzzy partition matrix $U = (\mu_{ij})_{n\times c}$ (introduced in Section 2.1.4), represents the membership degree of data point i to its cluster centre j. The higher the value of μ_{ij}, the stronger the data point i belongs to cluster j.

For fuzzy clustering, several validity indices have been defined in literature and a very simple distinction can be made separating those involving only the membership values from those also involving the data set. Among the first group, the Partition Coefficient, the Partition Entropy Coefficient and the Proportion Exponent were considered. All the other indices analysed subsequently will be involving in their definition the data set too.

Partition Coefficient and Partition Entropy Coefficient

The **Partition Coefficient (PC)** was proposed by Bezdek *et al.* in 1984 [13] and was defined as:

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^2 \tag{2.14}$$

where n is the number of data points and c the number of clusters. The Partition Coefficient takes values in the range $[\frac{1}{c}, 1]$. When $PC = 1/c$, the situation where all data points have an equal membership to all cluster centres occurs, indicating that the clustering is the most fuzzy. Furthermore, a value close to $1/c$ indicates that there is no clustering tendency in the considered dataset or the clustering algorithm failed to reveal it. When $PC = 1$, all clusters have well-defined borders, which means that each data point has a membership to its cluster centre of one and the algorithm is actually the K-means one. Therefore, as the clustering quality increases, the value of PC also increases [74].

The **Partition Entropy Coefficient (PE)** was also proposed in [13] and was defined as:

$$PE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}\cdot\log_a(\mu_{ij}) \tag{2.15}$$

where the logarithmic base a is in $(1, +\infty)$ and $PE \in [0, \log_a(c)]$. Similarly to PC, there are two extreme situations corresponding to the extreme values of PE. When $PE = 0$,

the clusters are well separated and a similar situation to $PC = 1$ occurs. In contrast, when $PE = \log_a(c)$, the clustering is fuzzier and the adopted fuzzy algorithm is unable to extract the clustering structure. Thus, as the clustering quality increases, the value of PE decreases [74, 175].

According to Sun *et al.*, these two indices are easy to compute. They are also useful when the data contains only a small number of well-separated clusters. However, there is a lack of direct connection to the geometrical properties of the data [162]. Further limitations of PC and PE have been pointed out by Halkidi and colleagues in [74]:

1. they are monotonously dependent on the number of clusters. Thus, a significant increase (for PC) or decrease (for PE) is visible when the number of clusters c increases.

2. They are too sensitive to the fuzzifier, m. More specifically, as $m \to 1$, the indices give the same values for all values of c.

3. They lack direct connection to the geometry of the data, since they do not use the data itself.

The Proportion Exponent Validity Index

According to Windham [187], the most important question associated with the use of a clustering algorithm is simply how well has it identified structure that is present in the data. Windham defined a function which assigns to a collection of membership functions a real number called the proportion exponent. It is assumed that the output of a fuzzy clustering algorithm includes a $c \times n$ matrix, $U = (u_{ij})$, where c is the number of clusters identified and n is the number of data points and $2 \leq c \leq n$. Each row of U is associated with a particular cluster and each column is the membership function of a particular data point. For each $j = 1, \ldots, n$ define $\mu_j = \max_i(u_{ij})$ and I_j to be the greatest integer in $1/\mu_j$.

The **proportion exponent** of U, $P(U)$, is defined in [187] by

$$P(U) = -\log_2\left(\prod_{j=1}^{n}\left(\sum_{k=1}^{I_j}(-1)^{k+1}\binom{c}{k}(1-k\mu_j)^{c-1}\right)\right). \tag{2.16}$$

To justify the negative logarithm, Windham stated that it spreads the values of the functional over a much wider range, particularly for proportions near zero. It also implies that large values for the proportion exponent indicate that the algorithm has worked well [187].

It is important to note that the evaluation of the proportion exponent index does not involve the data or the algorithm used to partition them and its maximum implies the optimal partition but without knowing what maximum is a statistically significant maximum. Moreover, $0 \leq P(U) < \infty$, since the $[0,1]$ values of the argument in (2.16) explode to $[0,\infty)$ due to the logarithm. Specifically, $P = 0$ when and only when $U = [1/c]$, while $P \to \infty$ when any column of U is crisp [12].

The Separation Index

This index identifies unique cluster structure with well-defined properties that depend on the data and a measure of distance. It seeks compact and separated (CS) clusters, but it rather seems computationally infeasible for big data sets since a distance matrix between all the data membership values has to be calculated. It also presupposes that a hard partition is derived from the fuzzy one. The **separation index** D_1 was firstly introduced by Dunn [44] and defined as

$$D_1 = \min_{1\leq i\leq c}\left\{\min_{i+1\leq j\leq c-1}\left\{\frac{\mathrm{dis}(u_i,u_j)}{\max_{1\leq k\leq c}\{\mathrm{dia}(u_k)\}}\right\}\right\},$$

where the diameter of the subset u_k is given by

$$\mathrm{dia}(u_k) = \max_{X_i,X_j\in u_k} d(X_i,X_j),$$

the distance of two subsets u_i and u_j is

$$\mathrm{dis}(u_i,u_j) = \min_{X_I \in u_i, X_j \in u_j} d(X_i,X_j)$$

and d is any metric induced by an inner product on the data definition space. The CS clustering is to be found by solving

$$\max_{2\leq c\leq n}\{\max_{\Omega_c} D_1\},$$

where Ω_c denotes the optimality candidates at fixed c [191]. It has been proved [12, 45] that a hard c-partition of the data set contains c compact, separate clusters if and only if $D_1 > 1$. Furthermore, Dunn proved that if there is a partition U such that $D_1 > 1$, that partition U is unique [45]. This result shows that CS partitions are distinguished by uniqueness whenever they exist. Since there is at most one such hard clustering of the data set at each c, a validity strategy based on maximising D_1 over all partitions is well defined via this unique limit. The proof itself depends on a simple observation: since the data set X is fixed, any pair of distinct c-partitions of X must intersect [12].

Xie Beni Validity Index

In 1991, Xie and Beni presented a fuzzy validity criterion based on a validity function which identifies overall compact and separate fuzzy c-partitions without assumptions on the number of substructures inherent in the data [191]. This function depends on the data set, geometric distance measure, distance between cluster centroids, and more importantly on the fuzzy partition generated by any fuzzy algorithm used. The function is mathematically justified via its relationship to the hard clustering validity function, the separation index, just defined above.

In order to define their validity index, Xie and Beni introduced several definitions [191]. Consider a fuzzy c-partition of the data set $X = \{X_j; j = 1,2,\ldots,n\}$ with $V_i (i = 1,2,\ldots,c)$ the centroid of each cluster and $\mu_{ij} (i = 1,2,\ldots,c, j = 1,2,\ldots,n)$ as the fuzzy

membership of data point j (also called vector j) belonging to class i.

Definition 2.1 *$d_{ij} = \mu_{ij}\|X_j - V_i\|$ is called the fuzzy deviation of X_j from class i. $\|.\|$ is the usual Euclidean norm and d_{ij} is just the Euclidean distance between X_j and V_i weighted by the fuzzy membership function.*

Definition 2.2 *$n_i = \sum_{x_j} \mu_{ij}$ is the fuzzy number of vectors in or fuzzy cardinality of class i. Summing up the n_is over x_i the total number of observations n is obtained.*

Definition 2.3 *The variation of class i, denoted by σ_i and defined as $\sigma_i = \sum_{x_j} (d_{ij})^2 = (d_{i1})^2 + (d_{i2})^2 + \ldots + (d_{in})^2$, is the summation of the squares of fuzzy deviation of each data point for each class i. The total variation of the data set X with respect to the fuzzy c-partition is defined as $\sigma = \sum_{x_i} \sigma_i$.*

It is important to note that both σ_i and σ depend on the fuzzy c-partition. A better c-partition should result in smaller σ.

Definition 2.4 *The ratio, denoted by π, of the total variation to the size of the data set, that is, $\pi = (\sigma/n)$, is called the compactness of the fuzzy c-partition of the data set.*

The value π measures how compact each class is. The more compact the classes are, the smaller π is. π is independent of the number of data points. For a given data set, a smaller π indicates that a partition with more compact clusters has been reached, thus indicating a better partition.

Definition 2.5 *The quantity $\pi_i = (\sigma_i/n_i)$ is called the compactness of class i.*

Definition 2.6 *The separation of the fuzzy c-partition is defined by $s = (d_{min})^2$, where d_{min} is the minimum distance between cluster centres, i.e., $d_{min} = \min_{i,j} \|V_i - V_j\|$.*

A large value of s indicates that all clusters are separated.

Definition 2.7 (Xie and Beni validity index) *The Xie and Beni validity index S, also called 'the compactness and separation validity function', is defined as the ratio between the compactness π and the separation s, i.e., $S = \pi/s$.*

A small value of S indicates that all the clusters in the partition under consideration are overall compact, and separate to each others. Thus, the fuzzy c-partition of X which minimises the value of S is considered as the best one for the analysis.

The Xie and Beni validity index can also be written in the following way:

$$S_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|V_i - X_j\|^2}{n\min_{i,j}\|V_i - V_j\|^2}, \tag{2.17}$$

and it can be easily seen that the definition of S is independent of the algorithm used to obtain μ_{ij} [191].

Gath and Geva Validity Index

Gath and Geva introduced three main criteria for comparing and finding optimal partitions based on the heuristics that a better clustering assumes i) clear separation between the clusters, ii) minimal volume of the clusters and iii) maximal number of data points concentrated in the vicinity of the cluster centroids. The performance measures were based on criteria for hypervolume and density [67]. The fuzzy hypervolume, F_{HV} was defined as

$$F_{HV} = \sum_{j=1}^{c}[det(F_j)]^{1/2},$$

where $F_j = \frac{\sum_{i=1}^{n}u_{ij}(x_i - v_j)(x_i - v_j)'}{\sum_{i=1}^{n}u_{ij}}$, for the case when the fuzzifier is 2, c is the number of clusters, n the total number of data points and u_{ij} is the degrees of membership of x_i in the jth cluster.

The second criterion, the average partition density D_{PA}, was calculated from

$$D_{PA} = \frac{1}{c}\sum_{j=1}^{c}\frac{S_j}{[det(F_j)]^{1/2}},$$

where $S_j = \sum_{i=1}^{n}u_{ij}$.

Moreover, Gath and Geva also defined the partition density, which expresses the general partition density according to the physical definition of density and was calculated

by:

$$P_D = \frac{S}{F_{HV}},$$

where $S = \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}$.

The hypervolume criterion is related to the within-cluster scatter, but due to its fuzzy characteristics the F_{HV} is not a monotone function of c (number of clusters). An optimal partition in the data would be the one for which F_{HV} reaches its minimum.

Rezaee *et al.* Validity Index

Rezaee *et al.* [148] introduced a new cluster validity index, which assessed the separation between clusters and the cohesion within clusters, which were generated by the fuzzy c-means (FCM) algorithm. They noted that a reliable validation functional for the FCM must consider both the compactness and the separation of a fuzzy c-partition because if only the compactness requirement were considered, the best partition would be obtained when each data point were considered as a separate cluster. On the other hand, if only the optimal separation between clusters were considered, the best partition would be the data itself; the distance between a cluster (the total data set in this case) and itself is zero.

In order to avoid these situations they designed the 'Compose Within and Between scattering' (V_{CWB}) index. It was defined by combining the average of the scattering (variation) within the c clusters and the total scattering (separation) between the clusters. In particular, several definitions were needed before defining the index itself. Assume a data set $X = \{x_1, \ldots, x_n | x_i \in \mathbb{R}^p\}$ with a fuzzy partition in c clusters ($\mathbf{V} = \{v_i, \ldots, v_c\}$ indicating the cluster centres) and $\mathbf{U} = [u_{ij}(i = 1, \ldots, c;\ j = 1, \ldots, n)]$.

Definition 2.8 *The variance of the pattern set X is called* $\sigma(X) \in \mathbb{R}^p$ *with the value of the pth dimension defined as*

$$\sigma_x^p = \frac{1}{n} \sum_{j=1}^{n} (x_j^p - \bar{x}^p)^2,$$

where $\bar{x}^p$ *is the pth value of the grand mean of X* ($\bar{X} = \sum_{j=1}^{n} x_j / n, \forall x_j \in X$).

Definition 2.9 *The fuzzy variation of the cluster i is called* $\sigma(v_i) \in \mathbb{R}^p$ *with the pth value defined as*

$$\sigma_{v_i}^p = \frac{1}{n}\sum_{j=1}^{n} u_{ij}(x_j^p - v_i^p)^2.$$

Definition 2.10 *The average scattering for c clusters is defined as*

$$Scat(c) = \frac{\frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\|}{\|\sigma(X)\|},$$

where $\|x\| = (x' \cdot x)^{1/2}$.

Definition 2.11 *A distance functional* $Dis(c)$ *is defined as*

$$Dis(c) = \frac{D_{max}}{D_{min}}\sum_{j=1}^{c}\left(\sum_{z=1}^{c}\|v_j - v_z\|\right)^{-1},$$

where $D_{max} = \max\|v_i - v_k\|\ \forall i,k \in \{2,3,\ldots,c\}$ *is the maximum distance between the cluster centres. The* D_{min} *has the same definition as* D_{max}*, but for the minimum distance between the cluster prototypes.*

The validation index V_{CWB} proposed by Rezaee *et al.* is a combination of the last two definitions:

$$V_{CWB}(\mathbf{U},\mathbf{V}) = \alpha\,\mathrm{Scat}(c) + \mathrm{Dis}(c),$$

where α is a weighting factor equal to $\mathrm{Dis}(c_{\max})$ and it is needed because the two terms of V_{CWB} are of a different range [148].

In general, a small value for the scattering (Scat) within the clusters indicates a compact partition, whereas the separation (Dis) term increased with the number of clusters. For this reason, the authors proposed that a cluster number which minimises the validation index can be considered as an optimal value for the number of clusters present in the data [148].

A summary of several validity indices for fuzzy clustering is reported in Table 2.2.

Validity index	Functional description	Optimal cluster number
Partition coefficient	$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^2$	$\max PC$
Partition entropy	$PE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}\cdot\log_a(\mu_{ij})$	$\min PE$
Proportion exponent	$P(U) = -\log_2\left(\prod_{j=1}^{n}\left(\sum_{k=1}^{I_j}(-1)^{k+1}\binom{c}{k}(1-k\mu_j)^{c-1}\right)\right)$	$\max P(U)$
Separation index	$D_1 = \min_{1\le i\le c}\left\{\min_{i+1\le j\le c-1}\left\{\frac{\mathrm{dis}(u_i,u_j)}{\max_{1\le k\le c}\{\mathrm{dia}(u_k)\}}\right\}\right\}$	$\max_{2\le c\le n}\{\max_{\Omega_c} D_1\}$
Xie and Beni	$S_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\lvert\lvert V_i - X_j\rvert\rvert^m}{n\min_{i,j}\lVert V_i - V_j\rVert^2}$	$\min S_{XB}$
Gath and Geva (hypervolume)	$F_{HV} = \sum_{j=1}^{c}[det(F_j)]^{1/2}$	$\min F_{HV}$
Rezaee *et al.*	$V_{CWB}(\mathbf{U},\mathbf{V}) = \alpha\,\mathrm{Scat}(c) + \mathrm{Dis}(c)$	$\min V_{CWB}$

x_k is the k-th data point, v_i are cluster centres, c is the number of clusters, and u_{ik} is the membership value of data x_k of class c_i.

Table 2.2: Several validation functionals for fuzzy clustering [148]

2.2.4 Principal component analysis

When a clustering algorithm is applied over a dataset, results may be displayed through various different methods. One of the most widely used is the biplot, which shows information on both samples and variables emphasising the clusters by using different colours or symbols. Samples are displayed as points while variables are displayed using vectors. Biplots were firstly introduced by Gabriel in [63]. In order to represent multi-dimensional data, the original data space is transformed using Principal Component Analysis (PCA) [86], and then the points are displayed at their projected position on axes of the first kth principal components considered (usually two or three).

The main purpose of PCA is to reduce the dimensionality of a data set consisting of a large number of correlated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables [92]. The transformation is achieved by rotating the original axes to produce orthogonal axes that are

uncorrelated to each other. The rotation procedure is a linear transformation of the original dataset and, therefore, if all the variables are included in the rotation, then all information is preserved [86]. The first component can be described as the orthogonal projection into one dimension that maximises the variance of the projected points; removing this dimension and its associated variance, the second component is the projection into one dimension that maximises the variance of the projected new array of points, and so on [46].

Suppose that $\mathbf{x}$ is a vector of p random variables, and that the variances of the p random variables and the structure of the covariances or correlations between the p variables are of interest. Unless p is small, or the structure is very simple, it will often not be very helpful to simply look at the p variances and all of the $\frac{1}{2}p(p-1)$ correlations or covariances. An alternative approach is to look for a few ($\ll p$) derived variables that preserve most of the information given by these variances and correlations or covariances. Although PCA does not ignore covariances and correlations, it concentrates on variances. The first step is to look for a linear function $\alpha_1'\mathbf{x}$ of the elements of $\mathbf{x}$ having maximum variance, where α_1 is a vector of p constants $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$, and $'$ denotes transpose, so that

$$\alpha_1'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j.$$

Next, look for a linear function $\alpha_2'\mathbf{x}$, uncorrelated with $\alpha_1'\mathbf{x}$ having maximum variance, and so on, so that at the kth stage a linear function $\alpha_k'\mathbf{x}$ is found that has maximum variance subject to being uncorrelated with $\alpha_1'\mathbf{x}$, $\alpha_2'\mathbf{x}$, ..., $\alpha_{k-1}'\mathbf{x}$. The kth derived variable, $\alpha_k'\mathbf{x}$ is the kth PC. Up to p PCs could be found, but it is hoped, in general, that most of the variation in $\mathbf{x}$ will be accounted for by m PCs, where $m \ll p$. By transforming the original variables to PCs, a reduction in complexity is achieved [92].

The identification of PCs will be now described. Consider, for the moment, the case where the vector of random variables $\mathbf{x}$ has a known covariance matrix σ. This is the matrix whose (i,j)th element is the (known) covariance between the ith and jth elements of $\mathbf{x}$ when $i \neq j$, and the variance of the jth element of $\mathbf{x}$ when $i = j$. The more realistic

case, where σ is unknown, follows by replacing σ by a sample covariance matrix $\mathbf{S}$. It turns out that for $k = 1, 2, \ldots, p$, the kth PC is given by $z_k = \alpha_k' \mathbf{x}$ where α_k is an eigenvector of σ corresponding to its kth largest eigenvalue λ_k. Furthermore, if α_k is chosen to have unit length ($\alpha_k' \alpha_k = 1$), then $\text{var}(z_k) = \lambda_k$, where $\text{var}(z_k)$ denotes the variance of z_k. To derive the form of the PCs, consider first $\alpha_1' \mathbf{x}$; the vector α_1' maximises $\text{var}[\alpha_1' \mathbf{x}] = \alpha_1' \sigma \alpha_1$. It is clear that, as it stands, the maximum will not be achieved for finite α_1 so a normalization constraint must be imposed. The constraint used in the derivation is $\alpha_1' \alpha_1 = 1$, that is, the sum of squares of elements of α_1 equals 1. Other constraints, for example $\max_j \|\alpha_{1_j}\| = 1$, may be more useful in other circumstances. However, the use of constraints other than $\alpha_1' \alpha_1 = \textit{constant}$ in the derivation leads to a more difficult optimization problem, and it will produce a set of derived variables different from the PCs. To maximise $\alpha_1' \sigma \alpha_1$ subject to $\alpha_1' \alpha_1 = 1$, the standard approach is to use the technique of Lagrange multipliers. Maximise

$$\alpha_1' \sigma \alpha_1 - \lambda(\alpha_1' \alpha_1 - 1),$$

where λ is a Lagrange multiplier. Differentiating with respect to α_1 gives

$$2\sigma\alpha_1 - 2\lambda\alpha_1 = \mathbf{0},$$

or

$$[\sigma - \lambda \mathbf{I}_p]\alpha_1 = \mathbf{0},$$

which is an ordinary eigenproblem where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. Thus, λ is an eigenvalue of σ and α_1 is the corresponding eigenvector. To decide which of the p eigenvectors gives $\alpha_k' \mathbf{x}$ with maximum variance, note that the quantity to be maximised is

$$\alpha_1' \sigma \alpha_1 = \alpha_1' \lambda \alpha_1 = \lambda \alpha_1' \alpha_1 = \lambda$$

(in the last equality the constraint $\alpha_1' \alpha_1 = 1$ was used). Thus, λ should be as large as

possible and $\alpha_1'\mathbf{x}$ is the first principal component of $\mathbf{x}$. In general, the kth PC of $\mathbf{x}$ is $\alpha_k'\mathbf{x}$ and $\text{var}(\alpha_k'\mathbf{x}) = \lambda_k$, where λ_k is the kth largest eigenvalue of σ, and α_k is the corresponding eigenvector [92].

It is well recognised [86,92] that the first papers describing principal component analysis were written by Pearson in 1901 and by Hotelling in 1933 [83], with the latter describing the general procedure as it is known today. Hotelling's approach starts from the idea of factor analysis, but he prefers to call the "new independent variables [...] which determine the values of the original" ones [83] as 'components' to avoid confusion with the term 'factors' which has already been used in mathematics [92]. Hotelling chooses his 'components' so as to maximise their successive contributions to the total of the variances of the original variables, and calls the components that are derived in this way the 'principal components'. In [83] they are obtained using Lagrange multipliers and ending up with an eigenproblem. However, Hotelling's procedure differs from the one described above in three different aspects. First, he worked with a correlation, rather than covariance, matrix; second, he looked at the original variables expressed as linear functions of the components rather than components expressed in terms of the original variables; and third, he did not use matrix notation [92].

2.2.5 Agreement between classifications

In the contest of clustering validity, when more than a single clustering technique is used, a method for assessing the agreement between the different classifications returned is to measure this agreement through a particular index. This measurement may be also useful when comparing clustering results against external criteria and it may help in the definition of a consensus clustering among the techniques used.

In 1960, Cohen introduced the *kappa coefficient* κ as a statistical measure of inter-rater agreement for qualitative (categorical) items [30]. Until then, the most frequently used index had been percentage or proportion of agreement (p_o), which suffers in that it includes agreement which can be accounted for by chance. The κ index, instead, is

generally thought to be a more robust measure since it takes into account the agreement occurring by chance. Occasionally, the $k \times k$ table of joint categorical assignment frequencies had been treated as a contingency table, and the contingency coefficient, C, based on chi-square, χ^2, had been used as a measure of agreement. The defect of χ^2 in this context, and therefore of C, is that it indexes association and not necessarily agreement, which is the special kind of association of interest in reliability [31].

The original kappa index proposed by Cohen [30] was defined as

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{2.18}$$

where p_o is the observed proportion of agreement, and p_c is the proportion of agreement expected by chance. Cohen also presented large sample formula for the standard error of an observed κ

$$\sigma_\kappa \cong \sqrt{\frac{p_o(1-p_o)}{N(1-p_c)^2}} \tag{2.19}$$

used for setting confidence limits and performing two-samples hypothesis tests [31]. Kappa index yields negative values when there is less observed agreement than is expected by chance, zero when observed agreement can be (exactly) accounted for by chance, and unity when there is complete agreement.

Further developments of κ were presented by Cohen himself in 1968 when he introduced the *weighted kappa* index. These were motivated by studies in which it was the sense of the investigator that some disagreements in assignments, that is, some off-diagonal cells in the $k \times k$ matrix, were of greater gravity than others. The κ, instead, does not make such distinction, implicitly treating all disagreement cells equally.

The weighted kappa index κ_w was derived considering the following steps [31]. If the proportion of disagreement is defined as $q = 1 - p$, then $p = 1 - q$. Substituting $p_o = 1 - q_o$ and $p_c = 1 - q_c$ into Equation (2.18) and simplifying, it can be obtained

$$\kappa = \frac{q_c - q_o}{q_c} = 1 - \frac{q_o}{q_c} \tag{2.20}$$

an equation for κ expressed in terms of observed and chance disagreement. κ_w simply replaces q_o and q_c by proportions of weighted disagreement, q'_o and q'_c. To find the latter, each of the k^2 cells must have a disagreement weight, v_{ij}, where the ij subscript indexes the cell ($i, j = 1, \ldots, k$). These positive weights can be assigned by means of any judgment procedure set up to yield a ratio scale. It is convenient (even though not necessary) to assign zero to the 'perfect' agreement diagonal ($i = j$), that is, no disagreement. A weight which represents maximum disagreement (v_{max}) is assigned at the convenience of the investigator. For any set of v_{ij}, κ_w is invariant over any positive multiplicative transformation. The weights assigned are an integral part of how agreement is defined and therefore how it is measured with κ_w [31].

Proportions of weighted disagreement, observed by chance, are simply weighted functions over the k^2 cells of the p_{oij} and p_{cij}, respectively, namely

$$q'_o = \frac{\sum v_{ij} p_{oij}}{v_{max}} \tag{2.21}$$

$$q'_c = \frac{\sum v_{ij} p_{cij}}{v_{max}} \tag{2.22}$$

where the p_{oij} is the proportion of the joint judgments (N in number) observed in the ij cell, and the p_{cij} the proportion in the cell expected by chance. Weighted kappa is then given by

$$\kappa_w = 1 - \frac{q'_o}{q'_c}. \tag{2.23}$$

When (2.21) and (2.22) are substituted in (2.23), the v_{max} term drops out, and it simplifies to

$$\kappa_w = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{cij}}. \tag{2.24}$$

Like the 'unweighted' kappa index, κ_w is fully chance corrected.

An interesting aspect is the relationship between κ_w and κ. The kappa index is simply proportion of agreement (p_o) corrected for chance, and κ_w can readily be thought of as a generalisation of κ, proportion of weighted agreement corrected by chance. The

relationship may be more clearly understood if it is inverted: κ is a special case of κ_w. In κ_w one may differentially weight, using v_{ij}, the off-diagonal ($i \neq j$) cells, because it is meant to consider the various kinds of disagreement. For κ, the $k(k-1)$ off-diagonal cells representing disagreement are simply treated as if they all represented the same amount of disagreement. Thus, κ is the special case of κ_w where all disagreements are given the same weight [31].

The discussion about weighted kappa index so far have implicitly assigned equal weights to symmetric cells, that is, $v_{ij} = v_{ji}$. This is appropriate to the frame of reference of reliability, where the two sources of data are conceived as being of equal status, that is, as alternate forms. Some reflection suggests that the formal difference between reliability and validity lies in the contrast between equal status of the sources in the former and their differing status in validity, where one is a predictor and the other a criterion. When validity is being assessed, it may (but need not) be eminently reasonable for $v_{ij} \neq v_{ji}$ [31].

Another widely used measure to assess the agreement between classifications is the Rand index [146]. Given a set of n objects $S = \{O_1, \ldots, O_n\}$, let $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ represent two different partitions of the objects in S such that $\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$ and $u_i \cap u_{i'} = \varnothing = v_j \cap v_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq C$. One of the partition may be the external criterion and the other a clustering result or both of them may be clustering results. Let a be the number of pairs of objects that are placed in the same element in partition U and in the same element in partition V, and d be the number of pairs of objects in different elements in partitions U and V [194]. b and c equal the number of pairs of objects which are not co-clustered, as reported in Table 2.3.

U/V	co-clustered	not co-clustered
co-clustered	a	b
not co-clustered	c	d

Table 2.3: Contingency table for two partitions [57]

The Rand index [146] is defined simply as the fraction of agreement, i.e.

$$R(U,V) = (a+d)/\binom{n}{2}.$$

The Rand index lies between 0 and 1, as, by definition, it is normalised. When the two partitions are identical, the Rand index is 1 [194].

A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value. For this reason, Hubert and Arabie [85] defined the *adjusted Rand index* which corrects for this by assuming the general form

$$\frac{\text{index - expected index}}{\text{maximum index - expected index}}.$$

In this general form the index is bounded above by 1, and takes the value 0 when the index equals its expected value [193]. As for the Rand index, a higher adjusted Rand index means a higher correspondence between the two partitions. The adjusted Rand index proposed by Hubert and Arabie was recommended by Milligan and Cooper [127] as the measure of agreement even when comparing partitions having different number of clusters.

Filkov and Skiena [57] defined the complementary measure of the Rand index, and called it the *Rand distance*. The authors defined it as the frequency of pairwise disagreements between U and V

$$1 - R(U,V) = (b+c)/\binom{n}{2}.$$

2.2.6 Summary

At the beginning of this section, three general questions about cluster validation were presented. To answer the first one, 'How many clusters should be considered?', it is possible to run several times a clustering algorithm (each time with a different number of groups as input) and store, after each run, the value of the chosen index for that specific configuration. Then, the decision rule for a specific index can be used to select the best set of clusters for the problem under investigation.

A similar approach may be used to answer the second question ('How good are the partitions?'): as the majority of the indices are built considering the dispersion of data

points within and between clusters, a good grouping will be the one where clusters are compact and well separated, i.e. the one for which indices express a low dispersion of points within clusters and a high one between groups. This internal property makes the validity indices suited to the problem of selecting the proper number of groups in classical cluster analysis.

The problem of evaluating a better partition than the one found, which answers the last question reported at the beginning of this section, may be addressed considering the different scores for measuring the agreement between classifications. When more than a single clustering algorithm is used, the concordance between different groupings returned may be assessed resorting to agreement scores like Cohen's kappa or Rand index. These may be also helpful when clustering results need to be compared with external criteria (such as partitions obtained with completely different approaches).

All these approaches have been described in this section, analysing in detail several validity indices for both hard and fuzzy clustering methods, and presenting scores for assessing the concordance between different partitions.

2.3 Clustering for breast cancer data

After the seminal paper of Eisen and colleagues [47], proposing hierarchical clustering and the visual inspection of the dendrogram to discover unknown pattern of gene associations, the use of clustering has become more and more popular, especially for discovering profiles in cancer with respect to high-throughput genomic data. Important applications of the Eisen *et al.* method are the work of Bittner *et al.* [15] on clustering of cutaneous melanoma and the works of van't Veer *et al.* [170] and Perou *et al.* [137] on breast cancer. In particular, Perou *et al.* [137] characterised variation in gene expression patterns in a set of 65 surgical specimens of human breast tumours from 42 different individuals, using complementary DNA (cDNA) microarrays representing 8102 human genes. Sets of co-expressed genes were identified for which variation in messenger RNA levels could be

related to specific features of physiological variation. The tumours could be classified into subtypes distinguished by pervasive differences in their gene expression patterns [137]. Authors identified four molecular distinct breast cancer groups based on gene expression profiles using a hierarchical clustering algorithm: luminal epithelial/estrogen (ER) positive, HER2 positive, basal-like and normal breast-like.

A subsequent study by Sørlie *et al.* [158] extended Perou and colleagues' work by improving the breast cancer classification. In particular, a total of 85 cDNA microarray tissue samples representing 84 individuals were analysed to classify breast carcinomas and to correlate tumour characteristics to clinical outcome [158]. A novel finding compared to previous Perou's work [137], was that the luminal epithelial/estrogen receptor-positive group could be divided into at least two subgroups, each with a distinctive expression profile and different prognosis. Using a hierarchical clustering, the breast samples were separated into two large branches. One contained three subgroups previously defined [137], while the luminal/ER+ group was separated into three distinct subgroups, which were called luminal-A, B and C [158]. A difference in outcome was observed for tumours classified as luminal A versus luminal B + C. The latter group of tumours, according to Sørlie *et al.*, might represent a clinically distinct group with a different and worse disease course, in particular with respect to relapse [158].

In an additional work of 2003, Sørlie *et al.* refined their previous classification [158] by analysing a total of 122 breast tumours using hierarchical clustering based on patterns of expression of 534 "intrinsic" genes [159]. The genes used for classification were selected based on their similar expression levels between pairs of consecutive samples taken from the same tumour separated by 15 weeks of neoadjuvant treatment. The samples fell into five major subgroups, characterised by distinct variation in gene expression pattern, which were quite similar to the original ones and differed from them by the elimination of the luminal-C group [159]. These five breast cancer subtypes were also associated with significant difference in clinical outcome. The authors once again addressed the point that gene expression studies have emphasised a considerable diversity among breast tumours,

both biologically and clinically [137, 158, 159].

In a work published in 2003, Sotiriou and colleagues analysed gene expression patterns generated from cDNA microarrays in an unselected group of 99 node-negative and node-positive breast cancer patients [161]. Using unsupervised hierarchical cluster analysis on the 706 probe elements, selected as exhibiting high variability across all tumours, two main groups based on their ER status were identified, which correlated well with basal and luminal characteristics. This finding corroborated Perou's previous analysis [137] that the major factor discriminating the expression phenotype is ER status [161].

Analysing the results more in detail, it could be seen that the dendrogram further branched into smaller subgroups within the ER+ and ER- classes. Within the ER negative cluster were tumours with 'basal'-like expression characteristics (basal 1 subgroup and basal 2 subgroup). Furthermore, a subgroup distinct from the basal-like groups in the ER- subset was defined by a high rate of HER-2/neu overexpression. The ER+ subgroup showed differential expression of genes associated with ER activation. These were also genes that defined the ER positive cluster as having 'luminal' characteristics as defined in [137, 158]. Moreover, according to Sotiriou and colleagues [161], the ER+ cluster could have been further segregated into three smaller subclasses, namely luminals 1, 2 and 3, similar to the luminals A, B and C identified by Sørlie *et al.* [158].

The authors underlined the importance of their results, since a concordance with those of the earlier studies were shown despite the differences in patient populations, treatments used, and technology platforms used [161].

Whilst numerous studies have reported the above and other novel molecular subtypes, and assigned a prognostic significance to the proposed classes [22, 170, 184], they remain varied in their detailed classification [84]. However, the following breast cancer groups, identified in recent literature [158, 161], became a sort of gold-standard in cancer characterisation: three luminal-epithelial groups, a HER2 positive cluster, and one or two basal-like groups. On the other hand, more recent studies [72, 95] began to criticize these groups. In particular, Gusterson [72] addressed some of the negatives and positives gen-

erated by the term basal-like breast cancer, and questioned its existence as an entity. He argued that in such a rapidly advancing field like the breast cancer research, it is essential that initial and thought-provoking results do not become established as 'facts' without question. Furthermore, Gusterson stated that the identification of basal-like breast cancer on the basis of gene-expression profiling data has been misleading in some respect and that a clear, basic understanding of breast cancer biology is needed to fully interpret gene-expression profiling data, in particular to improve the treatment of patients with triple negative cancers [72].

It is also worth noting that the visual inspection of the dendrogram suggested by Eisen *et al.* [47] is an informal method to determine the number of clusters. Such a procedure was firstly criticized by Marriot [118], who stated that the hierarchical methods may detect a natural grouping if it is sufficiently obvious, but do not appear to be primarily directed to doing so. Also Goldstein and colleagues [70] questioned the hierarchical approach, as, they said, it can cause difficulty in assessing the validity of the grouping. In particular, in [70] authors aimed to highlight some of the issues that arise when hierarchical clustering techniques are used in the analysis of cDNA microarray data. To illustrate these issues, the work of Bittner *et al.* [15] on cutaneous melanoma was considered. It was reported that results of the various clustering algorithms yielded different sets of clusters that grouped together. For example, while average linkage cluster analysis yielded a cluster of 19 melanoma samples for Bittner *et al.* [15] data, a cluster analysis using complete linkage yielded a cluster of 22 melanomas, as opposed to 19 [70]. Another criticism raised by Goldstein and colleagues was that there is not a standard criterion or algorithm for choosing a cutoff point for a dendrogram. Rather, this choice is often made by visual inspection. These, undoubtedly, leads to a certain degree of subjectivity being introduced in the analysis. The last criticism introduced in [70] was related to the choice of samples to be included in the clustering. The authors stated that finding genes which are differentially expressed across arrays is a major aim of microarray analysis. Genes with very low expression levels in some samples but not in others could be expected to be the basis of

an unknown subclass of tumours; their removal from the data set may contribute to false negative results, and could also encourage false positive results [70].

An alternative approach to gene expression profiling is to use established robust laboratory technology, such as immunocytochemistry on formalin fixed paraffin embedded patient tumour samples. Protein biomarker panels with known relevance to breast cancer have been applied to large numbers of cases using tissue microarrays, exploring the existence and clinical significance of distinct breast cancer classes [1,5,21,39,40,87,100,115].

In particular, in [5] tumour biological profiles were explored on 633 archival tissue samples analysed by immunohistochemistry. Five validated markers were considered, namely, estrogen receptor (ER), progesterone receptors (PgR), Ki-67/MIB1 as a proliferation marker, HER2/NEU, and p53. For ER, PgR and HER2/NEU, the percentage of their expression values tended to distribute around the values of 0%, 10%, 25%, 50%, 75%, and 100%, and were therefore discretised on these values. The results obtained were analysed by three different clustering algorithms. A hierarchical agglomerative algorithm with Ward's generalised criterion and two non-hierarchical techniques, K-means and Partitioning Around Medoids, were applied on the dataset considered. Four different validity indices, applicable to both hierarchical and non-hierarchical techniques, were then used to select the different profiles (number of clusters) [5]. The best classification was obtained resorting to four clusters. In particular, three of them were identified according to low, intermediate and high ER/PgR levels. A further subdivision into two biologically distinct subtypes was determined by the presence/absence of HER2/NEU and of p53. As previously noted [158, 159], the cluster with high ER/PgR levels was characterised by a better prognosis and response to hormone therapy compared to that with the lowest ER/PR values. Notably, the cluster characterised by high HER2/NEU levels showed intermediate prognosis, but a rather poor response to hormone therapy [5]. Significant achievements of this study were the consistency on the number of identified clusters with results of similar studies [115] and a partial difference with earlier works [137, 140, 170], were a tendency to identify only two profiles using hierarchical clustering was evident [5]. However, the

authors were well aware that the suggested four-clusters solution did not imply that the correct number of tumour subtypes was truly four. Increasing the number of investigated markers might lead to an increase of the clusters number and to finer subdivisions.

Another study involving immunohistochemistry on tumour samples was the one of Abd El-Rehim *et al.* [1] in which five breast cancer classes were defined. A sixth group of only four cases was also identified but considered too small for further detailed assessment. Groups 1 and 2 contained cancer that were luminal epithelial cell and hormone receptor positive; two additional groups (3 and 6) were characterised by high c-erbB-2 positivity and negative or weak hormone receptors expression but showed differences in MUC1 and E-cadherin expression. The fifth group (group 5) was characterised by strong basal epithelial characteristics, p53 positivity, absent hormone receptors and weak to low luminal epithelial cytokeratin expression. The final group (group 4), consisting of only four tumours, was difficult to describe; however, it appeared to be characterised by a basal phenotype with negative hormone receptor expression and strong expression of c-erbB2, p53 and nuclear BRCA1. This dataset will be described more in detail in the next chapter.

Existing studies have also not addressed the stability of the proposed classifications across different case sets, assay methods and data analysis procedures. Such an issue appears of critical relevance considering the increase in the number of features involved in bionformatics analyses. Moreover, especially in breast cancer studies, it is important to resort to different clustering techniques rather than focusing the attention on a single one, as diverse unsupervised classification methods usually return different groupings. To combine the classifications returned by several algorithms, many approaches have been proposed in the past. In the following section, a literature review on consensus clustering will be presented.

2.4 Consensus clustering

From a methodological perspective, to deal with the stability of classifications and in particular of clustering techniques, several studies focused on the comparison and concordance among different clustering methods defining what is now known as 'consensus clustering'.

Monti and colleagues in 2003 [130] presented a new methodology of class discovery and clustering validation tailored to the task of analysing gene expression data. The method can best be thought to guide and assist in the use of any of a wide range of available clustering algorithms. They call the new methodology consensus clustering, and in conjunction with resampling techniques, it provides for a method to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters. The method can also be used to represent the consensus over multiple runs of a clustering algorithm with random restart (such as K-means, SOM, etc.), so as to account for its sensitivity to the initial conditions. Cluster analysis, which aims to discover distinct and non-overlapping sub-populations within a larger population, is of particular significance in the field of gene expression data analysis. One of the fundamental issues to be addressed when clustering data (and especially in gene expression data analysis) includes how to assign confidence to the selected number of clusters, as well as to the induced cluster assignments, since the clustering results are especially sensitive to noise and susceptible to over-fitting. The basic assumption of the method proposed by Monti *et al.* was the following: if the data represent a sample of items drawn from distinct sub-populations, and if a different sample drawn from the same sub-populations were to be observed, the induced cluster composition and number should not be radically different. Therefore, the more the attained clusters are robust to sampling variability, the more one can be confident that these clusters represent real structure. To this end, perturbations of the original data can be simulated by resampling techniques. The clustering algorithm of choice can then be applied to each of the perturbed data sets, and the agreement, or consensus, among the multiple runs can be assessed. Consensus clustering simply formalizes

this procedure.

In a work of Swift and colleagues [163], consensus clustering was used to improve confidence in gene-expression analysis as authors stated that microarray analysis using clustering algorithms can suffer from lack of inter-method consistency in assigning related gene-expression profiles to clusters. In this paper it was recognised that many different heuristic algorithms, from the representative statistical ones (like K-means, Hierarchical and PAM) to the artificial intelligence techniques (such as genetic algorithms, neural networks and simulated annealing), have been used for partitioning gene-expression data with notable success [47]. To assess gene-expression cluster consistency, the use of the weighted-kappa metric was analysed by Swift *et al.* This metric is generally used as a comparison between two data partitions as it rates the agreement between the classification decisions made by two or more observers (see Section 2.2.5 for details). In this case the two observers are the clustering methods. The weighted-kappa compares clusters to generate score within the range -1 (no concordance) to +1 (complete concordance). However, the weighted-kappa metric shows that, even for highly correlated gene-expression profiles no two clustering algorithms have complete agreement. Overall this emphasizes that no single analysis method will identify all patterns in the gene-expression data; therefore multiple analyses should be performed and compared. Robust clustering and consensus clustering were suggested by the authors to achieve this goal.

Filkov and Skiena, in 2003 [57], proposed a methodology for consensus clustering as an approach to integrating diverse sources of similarly clustered microarray data. They proposed to exploit the popularity of cluster analysis of biological data by integrating clusterings from existing data sets into a single representative clustering based on pairwise similarities of the clusterings. Under reasonable conditions, the consensus cluster should provide additional information to that of the union of individual data analyses. The goals of consensus clustering are to integrate multiple data sets for ease of inspection, and to eliminate the likely noise and incongruencies from the original classifications. An investigation of the use of consensus clustering for increasing the reliability

of microarray gene expression data was performed by Filkov and Skiena [57]. It was shown that the consensus clustering is a robust approach, even when derived from small numbers of independent observations. Before studying the consensus clustering, Filkov and Skiena introduced the comparison of clusters by resorting to the Rand index and by defining the Rand distance. These two measures are complementary, as already shown in Section 2.2.5. The un-normalised form of the Rand distance was accepted as measure of choice. The consensus clustering was introduced by the authors by assessing that a consensus set-partition should be representative of the given set partitions. In terms of similarity it should be close to all given ones, or in terms of distance, it must not be too far from any of them. One way to do this is to find a partition that minimises the distance to all the other partitions. So, given k different partitions, Filkov and Skiena named the sought one as the 'consensus partition'.

Another approach has been used by Kellam and colleagues [98] where robust clusters were identified by the implementation of a new algorithm called by the authors 'Clusterfusion'. 'Clusterfusion' takes the results of different clustering algorithms and generates a set of robust clusters based upon the consensus of the different results of each algorithm. Firstly, an agreement matrix was generated with each cell containing the number of agreements amongst methods for clustering together the two variables represented by the indexing row and column indices. This matrix was then used to cluster variables based upon their cluster agreement. In essence, the authors applied a clustering technique to the clustering results.

2.5 Model-based clustering

In previous sections, a review of different clustering algorithms was presented. All those techniques find clusters following a heuristic approach, i.e. by optimising some criteria that depend on a distance between case pairs or between pairs of centroids of case collections [123]. It is important to realise that there is another big group of clustering

algorithms, called model-based clustering, which are based on probability models and that offer a principled alternative to heuristic algorithms. In particular, model-based clustering assumes that data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. The Gaussian mixture model has been shown to be a powerful tool for many applications [9]. With the underlying probability model, the issues of selecting a good clustering method and determining the correct number of clusters are reduced to model selection problems in the probability framework [58]. This provides a great advantage over heuristic clustering algorithms, for which there is no established method to determine the number of clusters or the best clustering method [192]. Model-based clustering has recently gained widespread use both for continuous and discrete domains [9] mainly due to the fact that it allows one to identify clusters based on their shape and structure rather than on proximity between data points [123].

With respect of standard heuristic clustering, the model-based approach has a couple of advantages. Firstly, at each stage of hierarchical clustering, the splitting or merging is chosen so as to optimize some criterion. In model-based methods, instead, a maximum-likelihood criterion is used for merging groups [9]. Secondly, relocation methods move observations iteratively from one group to another, starting from an initial partition. The number of groups has to be specified in advance and typically does not change during the course of the iteration. The most common relocation method (k-means) reduces the within-group sums of squares. For clustering via mixture models, relocation techniques are usually based on the EM algorithm (see Section 2.5.1) [58].

In model-based clustering it is assumed that data are generated by a mixture of G underlying probability distributions in which each component represents a different group or cluster. Given n observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, let $f_k(\mathbf{x}_i|\theta_k)$ be the density of an observation $\mathbf{x}_i$ from the kth component, where θ_k are the corresponding parameters [58]. The likelihood for the mixture model is

$$\mathcal{L}_{MIX}(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_n|\mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{x}_i|\theta_k), \tag{2.25}$$

where τ_k is the probability that an observation belongs to the kth components ($\tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1$). These τ_k are also called mixing parameters [14]. In the Gaussian mixture model, i.e. when $f_k(\mathbf{x}_i|\theta_k)$ is multivariate normal, each component k is modeled by the multivariate normal distribution with parameters μ_k (mean vector) and σ_k (covariance matrix):

$$f_k(\mathbf{x}_i|\mu_k, \sigma_k) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \sigma_k^{-1}(\mathbf{x}_i - \mu_k)\}}{\sqrt{\det(2\pi\sigma_k)}}. \tag{2.26}$$

Clusters are ellipsoidal and centred at means μ_k. The covariances σ_k determine their other geometric characteristics, like shape, volume and orientation [192].

Banfield and Raftery [9] developed a model-based approach for clustering based on the parametrisation of the covariance matrix σ_k in terms of eigenvalue decomposition in the following form:

$$\sigma_k = D_k \Lambda_k D_k^T. \tag{2.27}$$

In (2.27), D_k is the orthogonal matrix of eigenvectors, Λ_k is a diagonal matrix with the eigenvalues of σ_k on the diagonal. The orientation of the principal components of σ_k is determined by D_k, while Λ_k specifies the size and shape of the density contours. Λ_k may be factorised as $\Lambda_k = \lambda_k A_k$, where λ_k is the first eigenvalue of σ_k, $A_k = diag\{\alpha_{1k}, \ldots, \alpha_{pk}\}$, and $1 = \alpha_{1k} \geq \alpha_{2k} \geq \ldots \geq \alpha_{pk} > 0$. Thus, the orientation of the kth cluster is determined by D_k, while λ_k and A_k determine, respectively, its size (in terms of volume occupied in the p-space) and its shape [9]. From this parametrisation, several previously developed criteria for maximising Equation (2.25) may be retrieved: $\sigma_k = \lambda I$ gives the sum of squares criterion [71], in which clusters are spherical and have equal volume [58]. Instead, when $\sigma_k = \sigma(k = 1, \ldots, G)$, the criterion of Friedman and Rubin [61] is given, where all clusters have the same shape, volume and orientation. Finally, the most general criterion of Scott and Symons [154], which assumes all the components to be different, is obtained when $\sigma_k = \lambda_k D_k A_k D_k^T$.

The model parameters may be estimated resorting to the Expectation Maximisation (EM) algorithm of Dempster, Laird and Rubin [35].

2.5.1 EM algorithm

The EM algorithm [35] is a general approach to find maximum likelihood solutions in the presence of incomplete (latent) data [58]. The goal of EM is to maximise the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients) [14]. In maximum likelihood estimation, one wishes to estimate the model parameter(s) for which the observed data are the most likely [16]. The algorithm will be motivated here by giving an informal treatment in the context of the Gaussian mixture model. When a maximum of the likelihood function is found, the condition that must be satisfied is the null value of the derivative of the natural logarithm of $\mathcal{L}_{MIX}(\theta,\tau|\mathbf{x})$ in (2.25) with respect to the means μ_k of the Gaussian components:

$$\sum_{i=1}^{n} \underbrace{\frac{\tau_k f_k(\mathbf{x}_i|\mu_k,\sigma_k)}{\sum_{j=1}^{G}\tau_j f_j(\mathbf{x}_i|\mu_j,\sigma_j)}}_{\gamma(z_{ik})} \sigma_k^{-1}(\mathbf{x}_i-\mu_k) = 0 \tag{2.28}$$

where Equation (2.26) has been used for the Gaussian distribution [14]. γ_{ik} represent the posterior probabilities (also called responsibilities) in contrast with the prior τ_ks . Assuming σ_k not to be singular, it can be multiplied by it and obtained

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{n}\gamma(z_{ik})\mathbf{x}_i,$$

where N_k is defined as $N_k = \sum_{i=1}^{n}\gamma(z_{ik})$ and can be interpreted as the effective number of points assigned to cluster k.

Setting to zero the derivative of $\ln \mathcal{L}_{MIX}(\theta,\tau|\mathbf{x})$ with respect to σ_k, and performing similar calculations as above, the following relation is obtained:

$$\sigma_k = \frac{1}{N_k}\sum_{i=1}^{n}\gamma(z_{ik})(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^T.$$

Finally, when maximising $\ln \mathcal{L}_{MIX}(\theta,\tau|\mathbf{x})$ with respect to the mixing coefficients, the constrains underlying τ_k definition have to be taken into account. With a little bit of algebra

(see [14] for details) it can be obtained

$$\tau_k = \frac{N_k}{n}.$$

The results so far suggest an iterative scheme for finding a solution to the maximum likelihood problem. At the beginning some initial values for the means, covariances and mixing coefficients have to be chosen. Then an alternation between the following two steps (E and M) is to be performed: in the *expectation* step, or E step, the current values for the parameters are used to evaluate the responsibilities (posterior probabilities) γ. Then, in the *maximisation* step, or M step, these probabilities are used to re-estimate the means, covariances and mixing coefficient using the above results. It was shown [14] that each update to the parameters resulting from the E and M steps increases the log likelihood function and that the algorithm converges when the change in the log likelihood function, or alternatively in the parameters, falls below some threshold. The EM algorithm procedure may be described as follows:

Initialise the means μ_k, covariances σ_k and mixing coefficients τ_k, and evaluate the initial value of the log likelihood.

Repeat

E step. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{ik}) = \frac{\tau_k f_k(\mathbf{x}_i|\mu_k,\sigma_k)}{\sum_{j=1}^{G}\tau_j f_j(\mathbf{x}_i|\mu_j,\sigma_j)}.$$

M step. Re-estimate the parameter using the current responsibilities

$$\begin{aligned}
\mu_k^{\text{new}} &= \frac{1}{N_k}\sum_{i=1}^{n}\gamma(z_{ik})\mathbf{x}_i \\
\sigma_k^{\text{new}} &= \frac{1}{N_k}\sum_{i=1}^{n}\gamma(z_{ik})(\mathbf{x}_i-\mu_k^{\text{new}})(\mathbf{x}_i-\mu_k^{\text{new}})^T \\
\tau_k^{\text{new}} &= \frac{N_k}{n}
\end{aligned}$$

where

$$N_k = \sum_{i=1}^{n} \gamma(z_{ik}).$$

Evaluate the log likelihood

until convergence criteria are satisfied.

Although convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration [16], it is worth noting that EM takes many more iterations to reach convergence compared with the K-means algorithm, and that each cycle requires significantly more computations. It is therefore common to run the K-means algorithm first in order to find a suitable initialisation for a Gaussian mixture model that is subsequently adapted using EM [14]. Moreover, the number of conditional probabilities associated with each observation is equal to the number of components in the mixture, so that the EM algorithm for clustering may not be practical for models with very large numbers of components. Finally, EM breaks down when the covariance matrix corresponding to one or more components becomes singular or nearly singular [58].

In literature, Banfield and Raftery [9] have addressed the problem of non-Gaussian distributions suggesting a practical framework for model-based non-Gaussian clustering. Fraley and Raftery [58] have considered the problem of determining the structure of clustered data, without any knowledge about the number of clusters or any information about their composition. Different models with varying geometric properties were obtained through Gaussian components and they were compared among each other. In Meilă and Heckerman [123], three basic algorithms for model-based clustering were compared, namely the EM method, a 'winner take all' version of the EM algorithm, and model-based agglomerative clustering. The Expectation Maximisation algorithm was found to significantly outperform the other two. Yeung *et al.* [192] applied model-based clustering to gene expression data, showing the advantages of this approach over heuristic clustering algorithms also by testing the Gaussian mixture assumption for different transformations of expression data.

More recently, Frey and Dueck explored the combination of model-based and heuristic approaches and proposed a new clustering algorithm, called Affinity Propagation [60], which will be described in Chapter 6. This method is part of the message-passing family of algorithms (like the max-product algorithm [103]), which are used for performing inference through local computation [14].

2.6 Supervised classification techniques

The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown [101]. If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled. By applying these unsupervised (clustering) algorithms, researchers hope to discover unknown, but useful, classes [89].

Inductive machine learning is the process of creating a classifier that can be used to generalize from new instances. The process of applying supervised machine learning to a real-world problem is described in Figure 2.4.

In the remaining part of this section, three different supervised classification algorithms will be reviewed, namely, the C4.5 decision tree, the Multilayer Perceptron Artificial Neural Network (ANN), and the naive Bayes classifier.

2.6.1 Decision trees

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. The problem of constructing optimal binary decision trees is an NP-complete problem and thus theoreticians have searched for

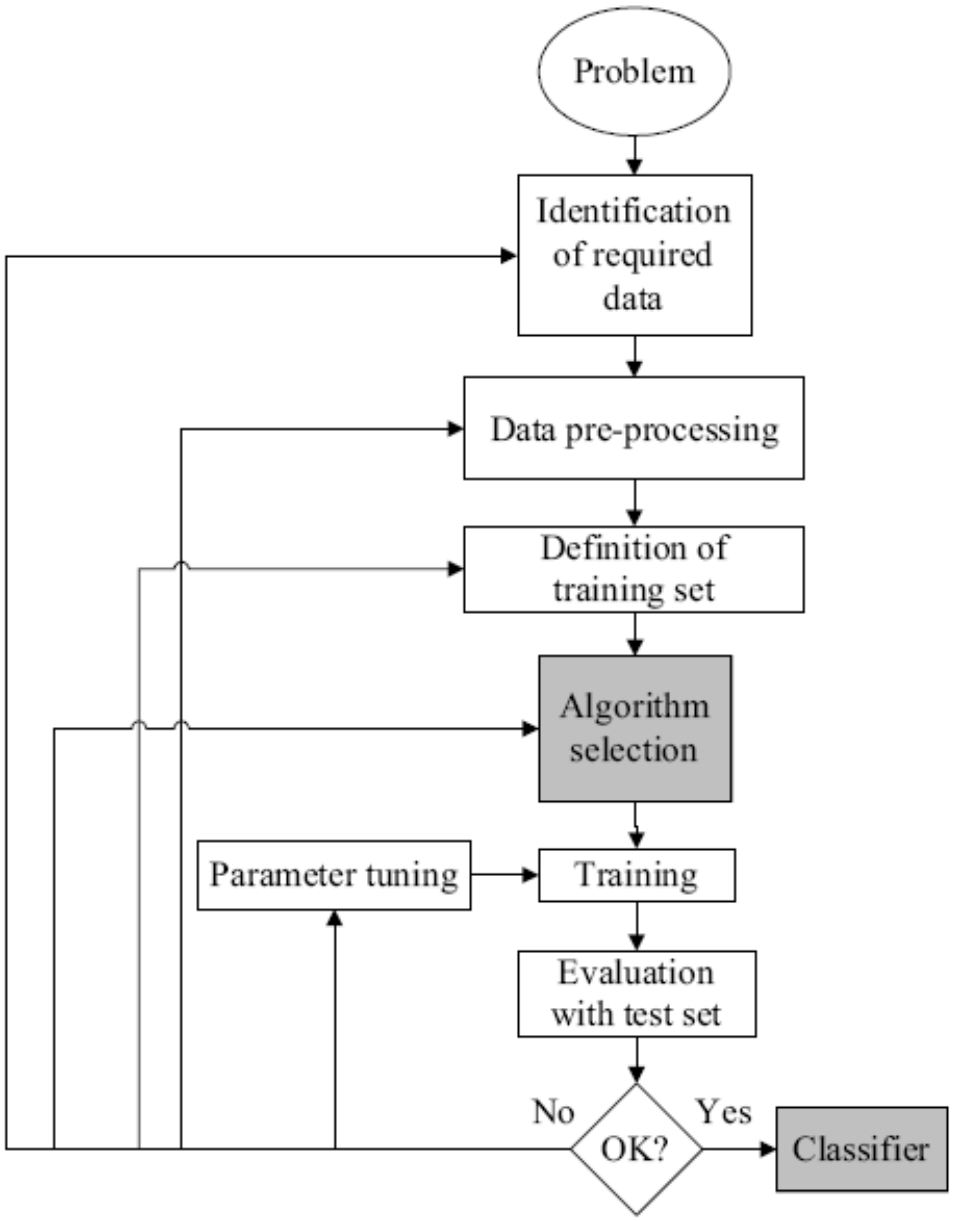


Figure 2.4: The process of supervised machine learning [101]

efficient heuristics for constructing near-optimal decision trees [101].

Perhaps the most well-know algorithm in literature for building decision trees is the C4.5 developed by Ross Quinlan [141]. C4.5 is an extension of Quinlan's earlier ID3 algorithm and it uses the concept of information gain to make a tree of classificatory decisions with respect to a previously chosen target classification [141]. Each attribute of the data can be used to make a decision that splits the data into smaller subsets. C4.5 examines the normalized information gain (difference in entropy) that results from choosing an

attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller sublists.

The output of the system is available as a symbolic rule base. The cases, described by any mixture of nominal and numeric properties, are scrutinized for patterns that allow the classes to be reliably discriminated. These patterns are then expressed as models, in the form of decision trees or sets of if-then rules, which can be used to classify new cases, with an emphasis on making the models understandable as well as accurate [141]. For real world databases the decision trees become huge and are always difficult to understand and interpret. However, decision trees tend to perform better when dealing with discrete/categorical features [101]. In general, it is often possible to prune a decision tree to obtain a simpler and more accurate tree [141].

2.6.2 Multilayer Perceptron ANN

A Multilayer Perceptron is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable by a hyperplane [80]. Typically, the network consists of a set of sensory units that constitute the *input layer*, one or more *hidden layers* of computation nodes, and an *output layer* of computation nodes. The number of nodes in the hidden layer must be large enough to form a decision region that is as complex as required by a given problem. Feed-forward ANNs allow signals to travel one way only, from input to output [101].

Multilayer Perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the *error back-propagation algorithm*. Basically, error back propagation learning consists of two passes through the different layers of the network: a forward pass and a backward pass. In the *forward pass*, an activity pattern (input vector) is applied to

the sensory nodes of the network, and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass the synaptic weights of the network are all *fixed*. During the *backward pass*, on the other hand, the synaptic weights are all *adjusted* in accordance with an error-correction rule. Specifically, the actual response of the network is subtracted from a desired (target) response to produce an *error signal*. This error signal is then propagated backward through the network, against the direction of synaptic connections – hence the name 'error back-propagation'. The synaptic weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense.

A Multilayer Perceptron has three distinctive characteristics:

1. The model of each neuron in the network includes a *nonlinear activation function*.

2. The network contains one or more layers of *hidden neurons* that are not part of the input or output of the network. These hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input patterns (vectors).

3. The network exhibits a high degree of *connectivity*, determined by the synapses of the network. A change in the connectivity of the network requires a change in the population of synaptic connections or their weights.

It is through the combination of these characteristics together with the ability to learn from experience through training that the Multilayer Perceptron derives its computing power [80].

2.6.3 Naive Bayes

A Bayesian classifier is a fast supervised classification technique which is suitable for large-scale prediction and classification tasks on complex and incomplete datasets. Naive Bayesian classification assumes that the variables are independent given the classes

[128]. It is also based on another common simplifying assumption: the values of numeric attributes are normally distributed within each class.

Let C be the random variable denoting the class of an instance and X be a vector of random variables denoting the observed attribute values. Let c be a particular class label and x represent a particular observed attribute value. According to the independence assumption, attributes $X_1 \ldots X_n$ are all conditionally independent of one another, given C. The value of this assumption is that it dramatically simplifies the representation of the conditional probability $P(X|C)$, and the problem of estimating it from the training data [129]. In fact, accurately estimating $P(X|C)$ typically requires many examples. To see why, just consider the number of parameters that must be estimated when C is boolean and X is a vector of n boolean attributes. In this case, the following set of parameters should be estimated:

$$\theta_{ij} \equiv P(X = x_i | C = c_j)$$

where the index i takes on 2^n possible values (one for each of the possible vector values of X), and j takes on 2 possible values. Therefore, approximately 2^{n+1} parameters need to be estimated. To calculate the exact number of required parameters, note for any fixed j, the sum over i of θ_{ij} must be one. Therefore, for any particular value c_j, and the 2^n possible values of x_i, only $2^n - 1$ independent parameters need to be computed. Given the two possible values for C one must estimate a total of $2(2^n - 1)$ such θ_{ij} parameters for learning Bayesian classifiers [129]. The Naive Bayes classifier, instead, reduces this complexity by making a conditional independence assumption that reduces the number of parameters to be estimated, when modeling $P(X|C)$, form the original $2(2^n - 1)$ to just $2n$. Moreover, to estimate $P(C|X)$, the training data can be used to learn estimates of $P(X|C)$ and $P(C)$. New X examples can then be classified using these estimated probability distributions, plus Bayes rule. This type of classifier is called a *generative* classifier, because the distribution $P(X|C)$ can be viewed as describing how to generate random instances X conditioned on the target attribute C [129].

If a test case x has to be classified, the probability of each class given the vector of

observed values for the predictive attributes may be obtained using the Bayes' theorem:

$$p(C = c|X = x) = \frac{p(C = c)p(X = x|C = c)}{p(X = x)} \tag{2.29}$$

and then predicting the most probable class. Because the event is a conjunction of attribute values assignments, and because of the attributes conditional independence assumption, the following equation may be written:

$$p(X = x|C = c) = \prod_i p(X_i = x_i|C = c)$$

which is quite simple to calculate for training and test data [91].

As previously mentioned, the second standard assumption of the classifier is that, within each class, the values of numeric attributes are normally distributed. One can represent such a distribution in terms of its mean and standard deviation, and the probability of an observed value from such estimates can be computed. For continuous attributes it can be written

$$p(X = x|C = c) = g(x;\mu_c,\sigma_c)\ , \quad \text{where} \tag{2.30}$$

$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.31}$$

is the probability density function for a normal (or Gaussian) distribution[1].

The major advantage of the naive Bayes classifier is its short computational time for training. In addition, since the model has the form of a product, it can be converted into a sum through the use of logarithms - with significant consequent computational advantages [101]. However, on many real-world datasets the latter condition is strongly violated. It might happen that, even in such situations, the naive Bayes classifier performs

[1]It is important to note that Equation (2.30) is not completely correct: the probability for a random variable of being exactly equal to any value is zero. Instead it is meant to consider a variable lying within a specific interval: $p(x \leq X \leq x+\varepsilon) = \int_x^{x+\varepsilon} g(x;\mu,\sigma)dx$. By the definition of derivative, $\lim_{\varepsilon\to 0} p(x \leq X \leq x+\varepsilon)/\varepsilon = g(x;\mu,\sigma)$. Thus, for very small values of ε, $p(X = x) \approx g(x;\mu,\sigma) \times \varepsilon$. The factor ε then appears in the numerator of Equation (2.29) for each class. They cancel out when the normalisation is performed, so Equation (2.30) may be used [91].

well, but one should always be aware that not all the hypotheses are satisfied.

According to John and Langley, methods for inducing probabilistic descriptions from training data have emerged as a major alternative to more established approaches to machine learning, such as decision-tree induction and neural networks. However, some of the most impressive results to date have come from the naive Bayes classifier, a much simpler – and much older – approach to probabilistic induction [91]. Despite the simplifying assumptions that underlie this classifier, experiments on real-world data have repeatedly shown it to be competitive with much more sophisticated induction algorithms. In [91] the assumption that data are generated by a single Gaussian distribution is abandoned because it is not always the best approximation. Authors suggest to investigate more general methods for density estimation, introducing what they call "Flexible Bayes", an extension of the naive Bayes classifier which uses a kernel density estimation. This method is very similar to the naive Bayes, but the density of each continuous variable is estimated averaging over a large set of kernels. The method performs well in domains that violated the normality assumption and, in general, this flexible Bayesian classifier generalizes better than the version that assumes a single Gaussian.

Bouckaert [17] also assesses that naive Bayes classifiers perform well over a wide range of classification problems, and, compared with more sophisticated schemes, they often perform better. He proposes a comparison of the three main methods for dealing with continuous variables in naive Bayes classifiers, namely the normal method, the kernel method and discretization. The normal method is the classical method that approximates the distribution of the continuous variable using a Gaussian distribution. The kernel method is the one cited above [91] which uses a non-parametric approximation. Finally, the discretization method [41] first discretizes the continuous variables into discrete ones, leaving a simpler problem without any continuous variable. In general, it is acknowledged that the normal method tends to perform worse than the other two methods. However, according to the simulations and experiments run by Bouckaert, none of the three methods systematically outperforms the others on all problems that were considered [17].

2.7 Similarities with Hyper-heuristics

According to [186] a hyper-heuristic is:

> "a heuristic search method that seeks to automate, often by the incorporation of machine learning techniques, the process of selecting, combining, generating or adapting several simpler heuristics (or components of such heuristics) to efficiently solve computational search problems. One of the motivations for studying hyper-heuristics is to build systems which can handle classes of problems rather than solving just one problem [19, 135, 149]."

There might be multiple heuristics from which one can choose for solving a problem, and each heuristic has its own strength and weakness. The idea is to automatically devise algorithms by combining the strength and compensating for the weakness of known heuristics [134]. In a typical hyper-heuristic framework there is a high-level methodology and a set of low-level heuristics (either constructive or perturbative heuristics). Given a problem instance, the high-level method selects which low-level heuristic should be applied at any given time, depending upon the current problem state, or search stage [149].

A similar approach to hyper-heuristics can be thought for the framework proposed in this thesis. By simply using the following correspondences it is possible to emphasise a broad analogy between the hyper-heuristics and the framework.

$$\begin{aligned} \text{Heuristic} &\longleftrightarrow \text{Clustering algorithm} \\ \text{Solving a problem} &\longleftrightarrow \text{Analysing a dataset} \\ \text{Hyper-heuristic} &\longleftrightarrow \text{This thesis framework} \end{aligned}$$

Using these correspondences, the first sentence of the second paragraph of this section could be re-written as follows: "There might be multiple clustering algorithms from which one can choose for analysing a dataset, and each algorithm has its own strength and weakness". The idea of the framework that will be presented in this thesis is to produce a single approach by combining the strength and compensating for the weakness of known clustering and classification methods.

2.8 Summary

In this chapter, a general overview of both the clustering techniques used within the literature and their application on breast cancer studies were provided. The purpose of clustering is to group objects so that they have the most similarity when belonging to the same cluster and the most dissimilarity when they are in different clusters. In particular, through the clustering process of gene expression data or tissue microarray, diverse breast cancer phenotypes have been defined in recent years. As one of the main objectives of this study is to emphasise the importance of applying different clustering methods on breast cancer studies rather than just a single one, particular consideration has been given to non-hierarchical algorithms and to the different consensus clustering approaches developed in the past. Moreover, it has been shown how a clear and well accepted definition of breast cancer groups is still far from being given.

In several clustering procedures it is required that the quality of the clustering results is verified. This might be achieved resorting to cluster validity measures. In this chapter, a particular section was dedicated to the identification of several cluster validity indices that have been proposed in literature to evaluate the partition results. Validation criteria are also used when the number of cluster is not know prior to commencing the analysis. In this way, different groupings may be analysed and the best one may be chosen, according to some validity optimisation rules.

To complement the overview on clustering, a general description of the model-based approach has been reported to point out other possible ways of dealing with grouping issues. A particular technique, called Affinity Propagation have been used in this study and a comparison between the CPU time needed for the algorithm computation and the time requested by K-means has been performed and is presented in Chapter 6.

To develop a general framework to elucidate core classes in a dataset, several supervised classification approaches have to be considered too. This chapter also identified several supervised learning algorithms, which have been used to build models of the class distribution labels in terms of predictor features and to predict the class assignment of

possible new objects. In Section 2.6 an overview of those techniques used in this thesis work was presented, together with a brief literature review on one of them.

In conclusion, three main aspects remain open for further investigation. Firstly, from the already published literature, it is clear that a range of clustering algorithms were used in breast cancer studies, leading to different definitions of cancer subtypes. However, it is also known that a perfect clustering method can not be defined, thus stressing the importance of a multi-techniques approach in this type of studies.

It also seems that a comprehensive and systematic comparison of clustering techniques on breast cancer data has not been carried out yet. This gap of knowledge can be filled by analysing the effect of a multi-techniques approach on the stability of results coming from clustering analysis for breast cancer data. Moreover, the comparison of different techniques will answer the first research question reported in Section 1.2.

The third aspect on which more focus is needed is related to consensus clustering. In this chapter, several approaches to address this issue were presented. For example, Monti *et al.* suggested using multiple runs of the same algorithm to form the consensus cluster [130]. Swift and colleagues proposed to measure the agreement between classifications using the kappa index [163]. The consensus clustering proposed by Filkov and Skiena was based on that partition which minimises the distance between all the others [57], while Kellam *et al.* suggested applying a clustering algorithm on the clustering results [98]. Despite all these different approaches for consensus clustering, an evaluation of the agreement between different methods applied to the same data combined with a heuristic combination of the groups derived by different techniques is still missing and leaves space for further investigation.

In the next chapter, an overview of the medical background relevant to this thesis will be presented in order to clarify the major aspects of the breast cancer disease, its diagnosis and the possible treatments. A description of how data were collected will be also given.

Chapter 3

Medical Background

This research is a multi-disciplinary work which involved both the School of Computer Science and the School of Molecular Medical Sciences at the University of Nottingham, UK. The close collaboration between the two Schools allowed the work to be conducted focusing on the aims and needs of both sides. The medical information presented in this chapter have been derived from joint publications (including several conference papers and journal papers).

The type of tumour that has been investigated in this study was the breast cancer. In order to examine whether a suspected patient has this type of cancer, tissue samples were collected using microarray technology and, therefore, this chapter will also describe the process of collecting, preparing and conducting microarray analysis on these samples.

3.1 Definition of breast cancer and treatments

Breast cancer is the most common cancer in the UK and the second most common worldwide. The latest statistics from Cancer Research UK website (2006) shows that each year more than 45,500 women and around 300 men are diagnosed with breast cancer in the UK. Just under 12,000 women and around 90 men die from breast cancer every year in the UK [181]. The ability to accurately identify the malignancy is crucial for prognosis and preparation of effective treatment. For breast cancer, some preoperative

imaging methodologies, such as x-ray mammography and ultrasound, can identify areas of tumour growth in the breast based on the identification of density changes within the tissue. The mammogram can detect small changes in breast tissue which may indicate cancers which are too small to be felt either by the woman herself or by a doctor [181].

Additionally, the diagnosis of breast cancer can often also be achieved by assessing the lymph nodes in the ipsilateral axilla (located on or affecting the same side of the axilla). The presence of metastasis (cancer spread from its original location) is an indicator for local disease recurrence and thus a method for identifying patients who are at high risk of developing a cancer variant that could spread throughout the body. The well-established procedure to access lymph node metastases is axillary lymph node dissection (ALND) [175].

The introduction of mammography screening programmes, together with a greater public awareness of breast cancer have meant that the majority of patients who do not have axillary lymph node metastases at presentation do not have to undergo ALND [147]. Intra-operative diagnosis has become increasingly important with the recent introduction of sentinel lymph node biopsy [168]. The sentinel node can be described as any lymph node that has a direct lymphatic connection to the tumour, and would be the first invaded by cancer spreading from the breast, as can be seen in Figure 3.1. Surgical studies have clearly shown that if cancer cannot be found in the sentinel lymph node, the chance of disease being found further down the chain of lymph nodes that drain the breast is negligible [168]. Therefore accurate analysis of the sentinel lymph node can alleviate the necessity to remove all suspected nodes present.

Surgery and radiotherapy are used to control local disease, and systemic treatments (chemotherapy and/or hormonal therapy) to combat frank or occult metastatic disease. Systemic treatments may also be administered up front as a primary treatment to reduce the size of the tumour prior to surgery.

Nearly all patients, whatever the stage of their disease, have some form of surgery. Other tests are carried out to assess the extent of the disease [181]. The main stages of

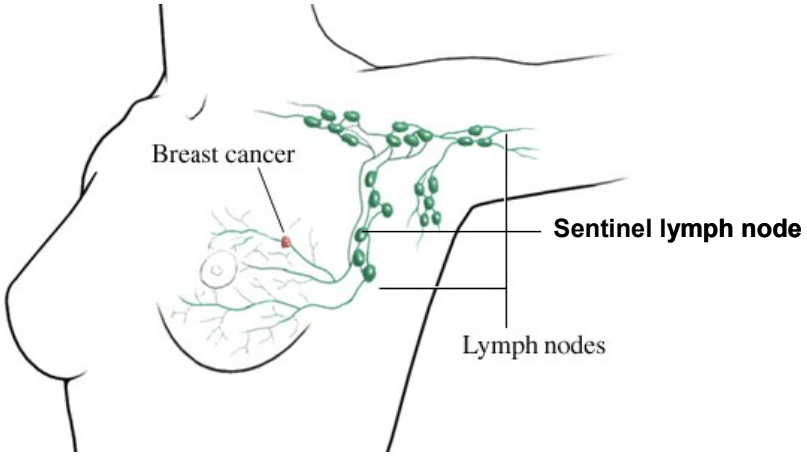


Figure 3.1: Typical location of lymph nodes that drain lymph from the breast

Stage	*Description*
Stage I	Tumour up to 2cm No lymph nodes affected No evidence of spread beyond the breast
Stage II	Tumour between 2cm and 5cm and/or; Lymph nodes in armpit affected No evidence of spread beyond armpit
Stage III	Tumour more than 5cm Lymph nodes in armpit affected No evidence of spread beyond armpit
Stage IV	Tumour of any size Lymph nodes in armpit often affected Cancer has spread to other parts of the body

Table 3.1: The main stages of breast cancer

invasive breast cancer are shown in Table 3.1.

A patient's treatment will depend upon a number of factors including the stage and grade of their tumour, hormone receptor (oestrogen and progesterone) status, menopausal status and general health. The standard treatment of lobular carcinoma in situ is surveillance, whereas ductal carcinoma in situ (DCIS) is often treated by complete local excision as there is a strong possibility that it will progress to invasive carcinoma [55].

Early breast cancer is potentially curable. Surgery is carried out to remove the tumour with an increasing trend towards more conservative surgery and reconstruction of the

breast.

The timing of surgery may be important: premenopausal women with early breast cancer seem to have a significantly better prognosis if their tumours are excised during the luteal phase I of the menstrual cycle [56]. During surgery, axillary lymph nodes are checked to see whether cancer has spread beyond the breast. New techniques of sampling lymph nodes may help to reduce the significant disability of lymphoedema of the arm. Some patients, for example young patients with large tumours, may receive chemotherapy before surgery (neo-adjuvant) to shrink the tumour, allowing more conservative surgery. Women who have oestrogen sensitive (ER positive) tumours receive some form of hormonal therapy to block the cancer-promoting effect of oestrogen [188]. Chemotherapy is usually given to women who have ER negative tumours although it may also be useful for some premenopausal ER positive patients.

Most patients do not present with advanced breast cancer. For those that do, some form of systemic treatment will be considered to control the cancer and improve quality of life [181].

In general, the main treatment options for breast cancer include [182]:

- **Surgery**

  The two most common operations for breast cancer are lumpectomy (surgery to remove the lump and some of the surrounding tissue) and mastectomy (removal of the whole breast). During most breast cancer operations, some lymph nodes are removed from the armpit (axillia). This is to look for cancer cells that have spread.

- **Radiotherapy**

  Quite often, women will have a course of radiotherapy starting two to four weeks after lumpectomy. This is to destroy any cancer cells that may still be present. Sometimes, women might also have radiotherapy after mastectomy. If the cancer has spread to other parts of the body, radiotherapy may be used to relieve symptoms such as bone pain.

- **Chemotherapy**

  Doctors often treat breast cancer with a combination of chemotherapy drugs. Women may receive chemotherapy before or after breast surgery. The doctor can also use chemotherapy to treat cancer that has come back.

- **Hormone therapy**

  The female hormone oestrogen is a major factor for the growth of many breast cancers. Hormone therapy lowers the amount of oestrogen in the blood, or blocks oestrogen from stimulating the cancer to grow. Tamoxifen is the most common hormone therapy used.

3.1.1 The Nottingham Prognostic Index

Lymph node status has been regarded for many years as the main indicator of prognosis [64]. It is a time-dependent prognostic factor – the longer the tumour has been growing the more likely it is to have spread to lymph nodes. Moreover, prognosis depends not only upon the presence of distant metastases but also upon their virulence. The virulence of a tumour depends on a number of intrinsic biological factors – some measurable, such as growth rate or response to hormone therapies, and some not yet so, such as invasiveness or power of tissue destruction [64]. To obtain a real power of prognostication, measures depending on both time-dependent factors and biological factors are needed. In recent years a new prognostic index was developed in order to assess how well treatments may work for a person with breast cancer and how long the person may live. This index was firstly introduced in 1982 [79] and was derived from a retrospective, multivariate study of nine factors in 387 patients with primary, operable breast cancer. In 1992 the Nottingham Prognostic Index (NPI) was applied to all of the first 1,629 patients in the series of operable breast cancer up to the age of 70 [64]. The index is defined as follows:

$$NPI\ Score = (0.2 \times size) + grade + stage$$

where *size* represents the tumour diameter in cm; *grade* means what the cancer cells look like under the microscope and ranges between 1 (low grade, slow growing) to 3 (high grade, faster growing); and *stage* represents the number of lymph nodes affected (1 if there are no lymph nodes affected, 2 if up to 3 glands are affected or 3 if more than 3 glands are affected). Among these three factors, size and lymph nodes stage are time-dependent, while histological grade is a biological characteristic. All the three factors have been proved to remain significant on multivariate analysis [64], so justifying the choice of considering them to define the NPI score.

Depending on the score, three groups are defined, namely a Good Prognostic Group (NPI < 3.4), a Moderate Prognostic Group ($3.4 \leq$ NPI ≤ 5.4) and a Poor Prognostic Group (NPI > 5.4). However, in other studies such as [111], five different groups of patients were be defined:

NPI Score	Prognostic Group
≤ 2.4	Excellent Prognostic Group (EPG)
2.5 - 3.4	Good Prognostic Group (GPG)
3.5 - 4.4	Moderate Prognostic Group 1 (MPG1)
4.5 - 5.4	Moderate Prognostic Group 2 (MPG2)
> 5.4	Poor Prognostic Group (PPG)

3.2 Instruments for breast cancer detection

When changes in breast shape or size are detected by a patient, it is advisable to consult a GP. The doctor will ask questions about the woman's medical history and about any risk factors she might have. The doctor will also carry out an examination of the breasts, armpits and neck and look for any lumps or suspicious changes. If necessary, a specialist may carry out further tests. These can include:

- **Mammograms**

  A mammogram is an X-ray of the breasts. Mammography is useful for finding early changes in the breast, when it may be difficult to feel a lump. It is not as helpful in younger women though. If a patient is under 35 years of age, it is more likely to

be suggested to have an ultrasound instead. Mammography can be painful because the breasts are put between two metal plates and a little pressure is applied. But most women describe this as mild to moderate discomfort, and it only lasts a few minutes. It is not harmful to the breasts [179].

- **Breast ultrasound scans**

  Breast ultrasound is painless and takes just a few minutes. Ultrasound uses sound waves to make a picture of the inside of the body. It is usually used for women under 35 whose breasts are too dense or solid to give a clear picture with mammograms. It is also used to see if a breast lump is solid, or if it contains fluid. A fluid filled lump is called a 'cyst' [179].

- **Tissue biopsy**

  A breast biopsy means taking a small sample of cells or tissue from your breast and looking at the sample under a microscope. A pathologist examines these samples and can see if they contain areas of cancer [179].

- **CT scan**

  CT scan (or CAT scan) stands for Computerised (Axial) Tomography scan. This just means a scan that takes a series of X-rays and uses a computer to put them together. The scan is painless. The CT machine takes pictures of the body from different angles and gives a series of cross sections or 'slices' through the part of the body being scanned. A very detailed picture of the inside of the body can be built up in this way. Together these cross sections give a very accurate picture of where a tumour is and how big it is. They also show how close major body organs are to the area that needs to be treated or operated on [178]. This kind of test is used for all kinds of cancers, not just for detecting breast tumours.

3.3 Microarrays

A microarray is a tool for analysing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. Microarrays allow scientists to analyse expression of many genes in a single experiment quickly and efficiently. They represent a major methodological advance and illustrate how the advent of new technologies provides powerful tools for researchers. Scientists are using microarray technology to try to understand fundamental aspects of growth and development as well as to explore the underlying genetic causes of many human diseases [27]. A microarray works by exploiting the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. By using an array containing many DNA samples, scientists can determine, in a single experiment, the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array [27]. With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

3.3.1 DNA microarray

DNA Microarrays are small, solid supports onto which the sequences from thousands of different genes are immobilized, or attached, at fixed locations. The supports themselves are usually glass microscope slides, but can also be silicon chips (in which case they are commonly known as *gene chip*) or nylon membranes. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. The DNA is printed, spotted, or actually synthesized directly onto the support [27].

It is important that the gene sequences in a microarray are attached to their support in an orderly or fixed way, because a researcher uses the location of each spot in the array to identify a particular gene sequence. The spots themselves can be DNA, cDNA, or oligonucleotides. An oligonucleotide is a short fragment of a single-stranded DNA

that is typically 5 to 50 nucleotides long. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system. DNA microarrays can be used to measure changes in expression levels or to detect single nucleotide polymorphisms (SNPs). Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analysing the data.

The use of a collection of distinct DNAs in arrays for expression profiling was first described in 1987 in [104], and the arrayed DNAs were used to identify genes whose expression is modulated by interferon. These early gene arrays were made by spotting cDNAs onto filter paper with a pin-spotting device. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, in a work of Schena and colleagues [153].

Arrays of DNA can be spatially arranged, as in the commonly known gene chip (also called genome chip, DNA chip or gene array, see Figure 3.2), or can be specific DNA sequences labelled such that they can be independently identified in solution. The traditional solid-phase array is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon biochip. The affixed DNA segments are known as 'probes' (although some sources use different terms such as 'reporters'). Thousands of them can be placed in known locations on a single DNA microarray. DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish that many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation [27].

Applications include gene expression profiling, which is the measurement of the activity (the expression) of thousands of genes at once, to create a global picture of cellular function. These profiles can, for example, distinguish between cells that are actively di-

Figure 3.2: Two DNA chips produced by Affymetrix

viding, or show how the cells react to a particular treatment. Many experiments of this sort measure an entire genome simultaneously, that is, every gene present in a particular cell. Furthermore, gene expression profiling has come into use as a way of defining, at the molecular level, the phenotypes of many kinds of tumours [159], thus leading to an explosion of molecular profiling studies [70].

A particular kind of microarray is the so-called 'two-channel microarray', which is typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) [155]. The two cDNA samples are mixed and hybridized to a single microarray (Figure 3.3) that is then scanned in a microarray scanner to visualize fluorescence after excitation with a laser beam of a defined wavelength [164].

3.3.2 Tissue microarray

Tissue microarrays (also TMAs) consist of paraffin blocks in which up to 1000 separate tissue cores are assembled in array fashion to allow multiplex histological analysis (Figure 3.4). These miniaturized collections of tissue spots result in a dramatic increase in throughput for in situ examination of gene status and gene expression from archival specimens [142].

The technique of tissue microarray was developed to address the issues of the cumber-

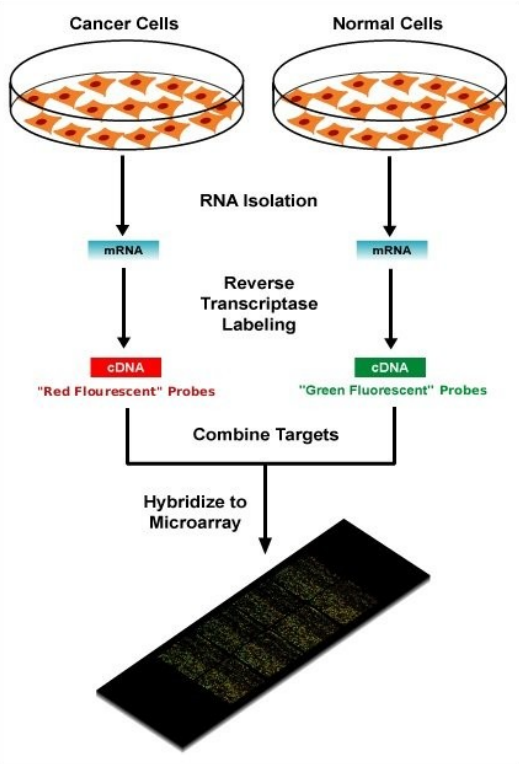


Figure 3.3: Diagram of typical dual-colour microarray experiment

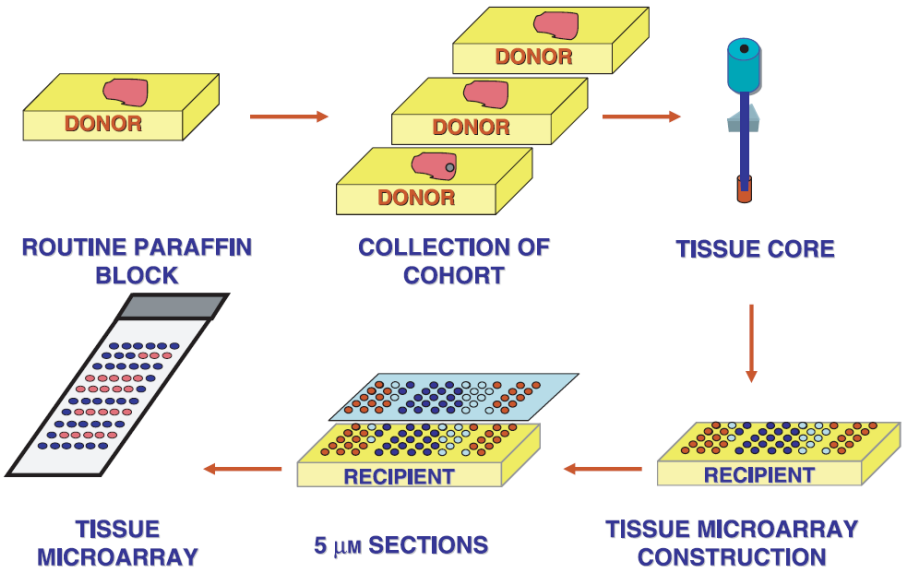


Figure 3.4: Process of tissue microarray construction

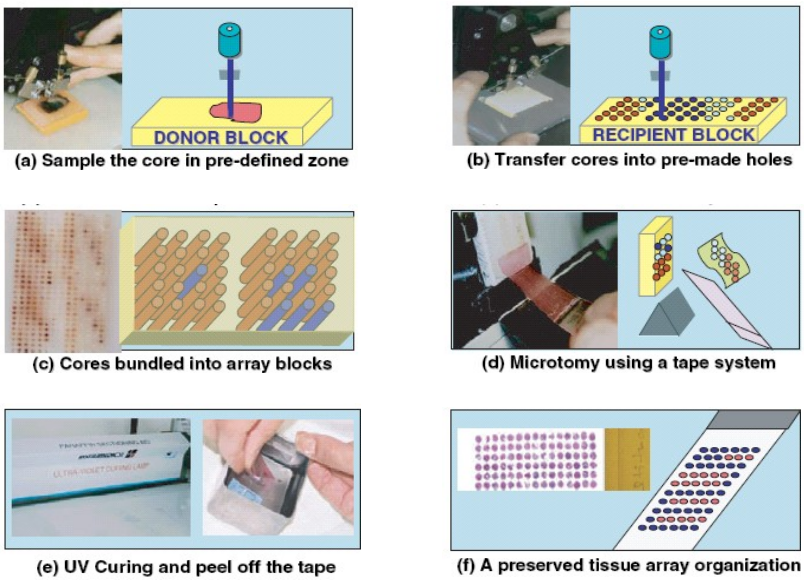


Figure 3.5: Construction of formalin-fixed paraffin-embedded tissue microarray

some nature of procedures, limited availability of diagnostic reagents and limited number of patients.

In the tissue microarray technique, an instrument built to create holes in the recipient block and for acquiring tissue core from the donor block is set up. The instrument consists of a thin wall stainless steel tube with an inner diameter of about 0.6 mm. The tissue cores are transferred into the recipient block through a solid stainless steel wire. As many as 1000 such tissue cores could be placed in one 45 $\times$ 20 mm recipient paraffin block. After the block construction is completed, 5 μm sections of the resulting tumour TMA block are cut and transferred onto a slide using an adhesive-coated tape. The complete procedure is described in detail in [142] and is shown in Figure 3.5.

Each microarray block can be cut into 100 – 500 sections, which can be subjected to independent tests. Tests commonly employed in tissue microarray include immunohistochemistry, and fluorescent in situ hybridization. Tissue microarrays are particularly useful in analysis of cancer samples [99].

An example of a 0.6mm core Tissue MicroArray Block is shown in Figure 3.6.

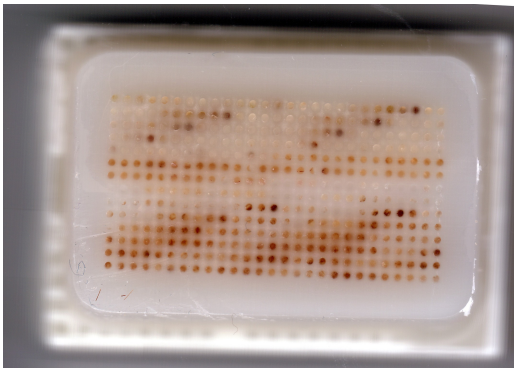


Figure 3.6: A 0.6mm core tissue microarray block

3.4 Data collection

3.4.1 Data pre-processing and Immunohistochemistry

For the analysis described in this thesis, breast cancer TMA were prepared as described in Section 3.3.2. Briefly, cores of 0.6 mm thickness were obtained from the most representative areas of the tumours then reembedded in microarray blocks. Each case was sampled twice; one core was obtained from the centre and the other from the periphery of the tumour. TMAs of 100 cases per block were made [1].

Immunohistochemistry (IHC) refers to the process of localizing proteins in cells of a tissue section exploiting the principle of antibodies binding specifically to antigens in biological tissues. Immunohistochemical staining is widely used in the diagnosis of cancer. Specific molecular markers are characteristic of particular cancer types. IHC is also widely used in basic research to understand the distribution and localization of biomarkers in different parts of a tissue [1].

Immunohistochemical staining for the sections was performed using the streptavidin-Biotin Complex method using a large panel of well-characterised commercially available tumour markers. To avoid loss or decline of immunoreactivity of tissue sections with increasing storage time, sections from TMA blocks were cut shortly prior to staining of each antibody.

The modified histochemical score (H-score) [121] was used because it includes a

semiquantitative assessment of both the intensity of staining and the percentage of positive cells. For the intensity, a score of 0 to 3, corresponding to negative, weak, moderate and strong positivity, was recorded. In addition, the percentage of positive cells at each intensity category was estimated. The H-score is calculated as follows [122]:

$$\begin{aligned}\text{H-score} = &(1\times\ \%\text{ of cells stained at intensisty category 1})\\ &+(2\times\ \%\text{ of cells stained at intensisty category 2})\\ &+(3\times\ \%\text{ of cells stained at intensisty category 3}).\end{aligned}$$

The range of possible scores is thus 0 to 300, where 300 equals 100% of tumour cells stained strongly [37]. H-score and similar semiquantitative scoring systems have been successfully used for TMA evaluation [1]. By using such a score, it was possible to explore rationalization of cases into biologically relevant groups depending on different levels of expression, which could not be obtained by using simpler scoring methods (e.g., positive vs. negative). Two cores were evaluated from each tumour. Each core was scored individually, then the mean of the two readings was calculated. If one core was uninformative (either lost or contained no tumour tissues), the overall score applied was that of the remaining core.

3.4.2 Assembling TMA and clinical data

For this research work, data were provided by pathologists and researchers of the School of Molecular Medical Sciences at Nottingham University Hospitals and University of Nottingham in form of several large datasets. In order to recompute several studies and to define a new version of the Nottingham Prognostic Index, there was the need to create a larger and up-to-date dataset full with all the available information. The process of assembling both clinical and biological information into a single resource with unique entries took time and was carried out through several stages using the SQL language. At the beginning, as all datasets were very heterogeneous (both in size and shape) and some of them even contained repeated variables or different variables' values for the same

patient, differences between data entries were pointed out, in order to select and keep the most recent data and to combine at least tables with the same set of patients.

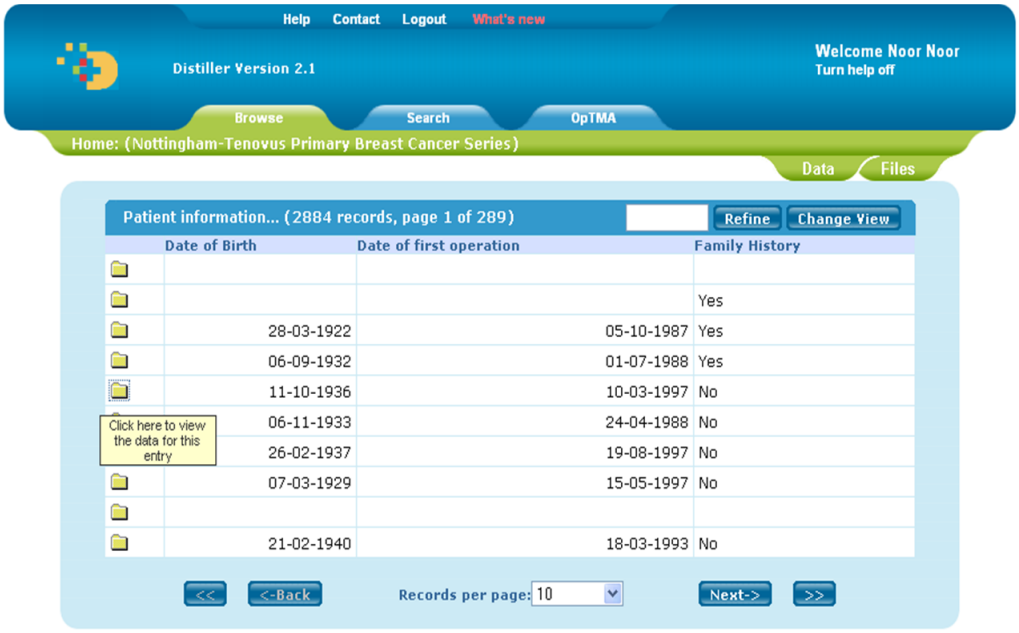
Subsequently, two resources containing respectively all the clinical information and all the biological ones were created by merging the previously 'cleaned' smaller tables with those ones provided in the meantime. Each final dataset contained approximately 230 different information for almost 2,450 patients (all women).

Finally, the available data was integrated with images taken by microscopes. This piece of work, as well as the maintenance and management of the data, were entrusted to 'Slidepath – Digital Slide Solution', a company dedicated to the delivery of world leading software for digital slide and life sciences informatics applications. In particular, they developed a new software called 'Distiller', which is a web based information management solution ideal for sharing files and data between distributed collaborators. Snapshots of the software are shown in Figure 3.7 and in Figure 3.8. Once logged in, a list of available markers and images, as well as the one of clinical information stored are displayed. By selecting the desired variables for the subset of patients of interest, a spreadsheet with all the information is created and downloaded on the local machine.

3.5 Summary

In this chapter, the medical background of this research work was presented, including the definition of breast cancer and the different instruments for its detection. A widely used prognostic measure, the Nottingham Prognostic Index was also introduced together with different kinds of breast cancer treatments. Microarray analysis techniques, which are used for interpreting the data generated from experiments on DNA, RNA, and protein microarrays were also presented. They allow researchers to investigate the expression state of a large number of genes in a single experiment. Finally, the processes followed for data collection were reported and the immunohistochemistry technique was presented.

The next chapter will present the application of several clustering techniques over



(a) List of patients

(b) Details for a specific patient

Figure 3.7: Snapshots of Distiller software

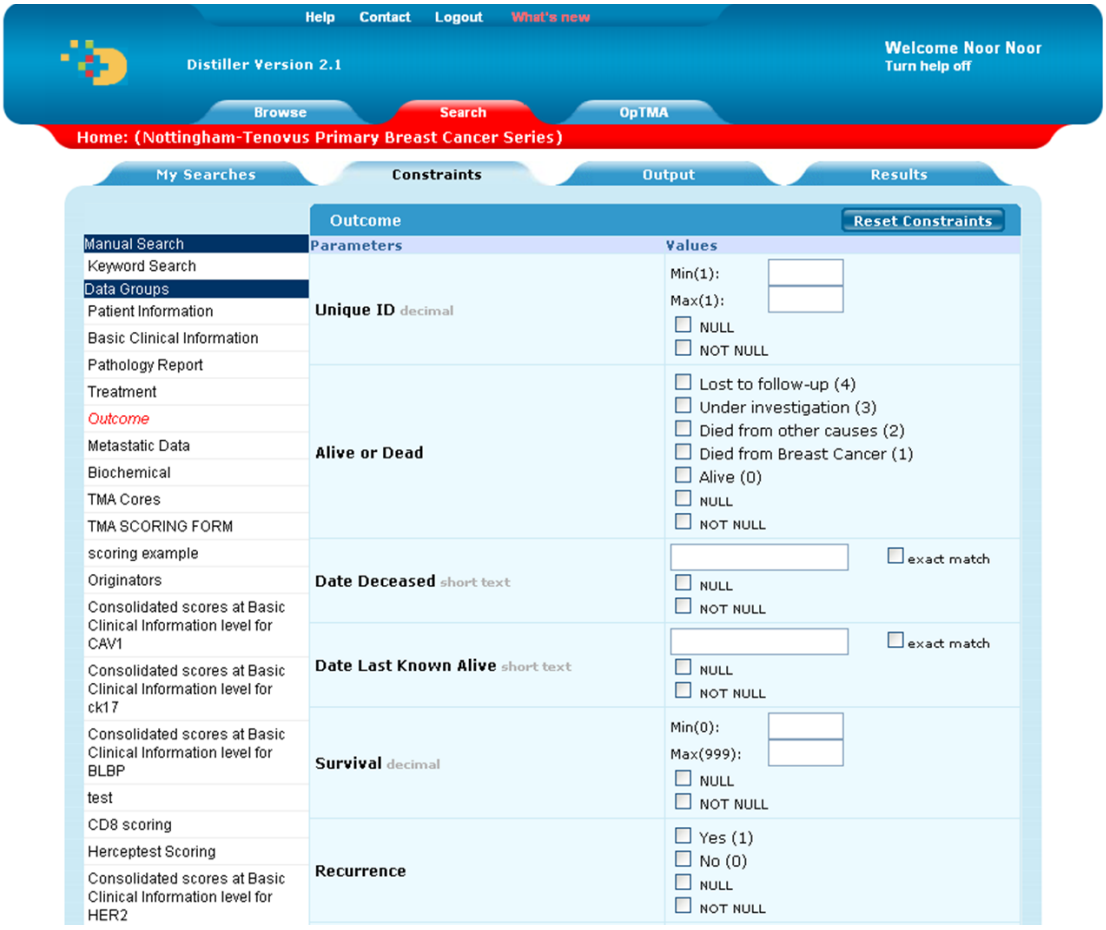


Figure 3.8: Possible constraints for the Search function of the Distiller software

breast cancer data and the use of an informal consensus clustering to categorise patients in representative classes.

Chapter 4

A Comparison of Different Clustering Techniques

4.1 Introduction

In this chapter, a comparison between different clustering algorithms applied over the same set of data will be described. In 2005, Abd El-Rehim *et al.* used a hierarchical approach to categorise breast cancer patients in six different groups which were relevant also in terms of clinical outcome [1]. However, it should be noted that one of these groups consisted of only four patients and, therefore, was difficult to characterise. In addition, as already pointed out in [70], the hierarchical approach has been criticized as it can cause difficulty in assessing the validity of the grouping. For these reasons various clustering algorithms were applied in an attempt to refine the previous classification.

It is worth noting that the idea of combining/comparing the results of different clustering algorithms is particularly important in order to evaluate the stability of the proposed classification. In this research work, the stability of six breast cancer classes was evaluated by comparing the different solutions provided by different algorithms. Concerning the standard problem of consensus clustering in which the label of classes is arbitrary, in this study a label was assigned using the six clusters characterised in the work of Abd

El-Rehim [1], as a reference for the description of the resulting groups.

4.2 Dataset description

A series of 1076 patients from the Nottingham Tenovus Primary Breast Carcinoma Series presenting with primary operable (stages I, II and III) invasive breast cancer between 1986-98 were used. Immunohistochemical reactivity for twenty-five proteins, with known relevance in breast cancer including those used in routine clinical practice, were previously determined using standard immunocytochemical techniques on tumour samples prepared as tissue microarrays [1]. Levels of immunohistochemical reactivity were determined by microscopical analysis using the modified H-score (values between 0-300), giving a semiquantitative assessment of both the intensity of staining and the percentage of positive cells. The complete list of variables used in this study is given in Table 4.2.

This is a well-characterised series [1] of patients who were treated according to standard clinical protocols. Patient management was based on tumour characteristics using Nottingham Prognostic Index (NPI) and hormone receptor status. Patients with an NPI score ≤ 3.4 received no adjuvant therapy, those with a NPI score > 3.4 received hormone therapy if oestrogen receptor (ER) positive or classical cyclophosphamide, methotrexate and 5-fluorouracil (CMF) if ER negative and fit enough to tolerate chemotherapy. Hormonal therapy was given to 420 patients (39%) and chemotherapy to 264 (24.5%). Data relating to survival was collated in a prospective manner for those patients presenting after 1989 only; including survival time, defined as the interval (in months) from the date of the primary treatment to the time of death. The overall survival was taken as the time (in months) from the date of the primary surgical treatment to the time of death. This study was approved by the *Nottingham Research Ethics Committee 2* under the title 'Development of a molecular genetic classification of breast cancer'.

Variable	Frequencies / Values
Age	
Min	18
Mean	53.4
Max	70
Tumour size	
Min	0.1
Mean	2.1
Max	10
Lymph node stage	
1	654
2	332
3	87
NA	3
Grade	
1	160
2	343
3	572
NA	1
NPI	
Min	2.1
Mean	4.3
Max	8.0
NA	4
Tumour type	
Invasive ductal / NST	649
Tubular mixed	171
Medullary	30
Lobular	112
Special types	46
Mixed NST & lobular	37
Mixed NST & special type	24
Miscellaneous	4
NA	3

Table 4.1: Distributions / frequencies of several variables in the dataset

4.3 Experiments

In this work, the starting point was the same data as in [1] and several unsupervised clustering techniques were applied in order to evaluate the stability of results coming from different clustering methods in terms of concordance among solutions. In the following (Section 4.3.1), the adopted clustering techniques will be described in detail, as well as the different methods used for characterising the classes created by the algorithms and the consensus among the different techniques (Section 4.3.4). The four-step methodology for elucidating core, stable classes (groups) of data from a complex, multidimensional dataset was as follows:

1. A variety of clustering algorithms were run on the data set.

2. Where appropriate, the most appropriate number of clusters was investigated by

Antibody, clone	Short Name	Dilution
Luminal phenotype		
CK 7/8 [clone CAM 5.2]	CK7/8	1:2
CK 18 [clone DC10]	CK18	1:50
CK 19 [clone BCK 108]	CK19	1:100
Basal Phenotype		
CK 5/6 [cloneD5/16134]	CK5/6	1:100
CK 14 [clone LL002]	CK14	1:100
SMA [clone 1A4]	Actin	1:2000
p63 ab-1 [clone 4A4]	p63	1:200
Hormone receptors		
ER [clone 1D5]	ER	1:80
PgR [clone PgR 636]	PgR	1:100
AR [clone F39.4.1]	AR	1:30
EGFR family members		
EGFR [clone EGFR.113]	EGFR	1:10
HER2/c-erbB-2	HER2	1:250
HER3/c-erbB-3 [clone RTJ1]	HER3	1:20
HER4/c-erbB-4 [clone HFR1]	HER4	6:4
Tumour suppressor genes		
p53 [clone DO7]	p53	1:50
nBRCA1 Ab-1 [clone MS110]	nBRCA1	1:150
Anti-FHIT [clone ZR44]	FHIT	1:600
Cell adhesion molecules		
Anti E-cad [clone HECD-1]	E-cad	1:10/20
Anti P-cad [clone 56]	P-cad	1:200
Mucins		
NCL-Muc-1 [clone Ma695]	MUC1	1:300
NCL-Muc-1 core [clone Ma552]	MUC1co	1:250
NCL muc2 [clone Ccp58]	MUC2	1:250
Apocrine differentiation		
Anti-GCDFP-15	GCDFP	1:30
Neuroendocrine differentiation		
Chromogranin A [clone DAK-A3]	Chromo	1:100
Synaptophysin [clone SY38]	Synapto	1:30

Table 4.2: Complete list of antibodies used and their dilutions

mean of cluster validity indices.

3. Concordance between clusters, assessed both visually and statistically, was used to guide the formation of stable ‘core’ classes of data.

4. A variety of methods were utilised to characterise the elucidated core classes.

Once these core classes were obtained, the clinical relevance of the corresponding patient groups was investigated by means of associations with related patient data. All the analysis described in this chapter, excluding the Adaptive Resonance Theory algorithm, was done using *R*, a free software environment for statistical computing and graphics [114].

4.3.1 Techniques considered

Five different algorithms were used for cluster analysis:

1. Hierarchical (HCA)
2. K-means (KM)
3. Partitioning around medoids (PAM)
4. Adaptive resonance theory (ART)
5. Fuzzy C-means (FCM)

Hierarchical clustering

The hierarchical clustering algorithm (HCA) begins with all data considered to be in a separate cluster. It then finds the pair of data with the minimum value of some specified distance metric; this pair is then assigned to one cluster. The process continues iteratively until all data are in the same (one) cluster. A conventional hierarchical clustering algorithm (HCA) was utilised in this research work, utilising Euclidean distance on the raw (unnormalised) data with all attributes equally weighted.

K-means clustering

The K-means technique aims to partition the data into K clusters such that the sum of squares from points to the assigned cluster centres is minimised. The algorithm repeatedly moves all cluster centres to the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre). The objective function minimised is:

$$J(V) = \sum_{j=1}^{k} \sum_{i=1}^{c_j} ||x_i - v_j||^2$$

where x_i is the i-th datum, v_j is the j-th cluster centre, k is the number of clusters, c_j is the number of data points in the cluster j and $||x_i - v_j||$ is the distance between x_i and v_j.

The j-th centre v_j can be calculated as:

$$v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i, \quad j = 1, ..., k.$$

K-means clustering is dependent on the initial setting of the cluster centres (which, in turn, determines the initial cluster assignments). Various techniques have been proposed for the initialisation of clusters [4], but for this study a fixed initialisation of the cluster centres obtained with hierarchical clustering was used. The Euclidean metric has been chosen to represent the distance between points and cluster centres and the maximum number of iterations was set to 100. The default algorithm of Hartigan and Wong [78] was used. The number of clusters is an explicit input parameter to the K-means algorithm.

Partitioning around medoids

The partitioning around medoids (PAM) algorithm (also known as the k-medoids algorithm) is a technique which attempts to minimise the distance between points labeled to be in a cluster and a point designated as the centre of that cluster. The main characteristics of this method are described in Section 2.1.3 of Chapter 2.

There are basically two ways of entering the data in PAM. The most common way is by means of a matrix of measurements values. The rows of this matrix represent the objects and the columns correspond to variables, which must be on an interval scale. Alternatively, the program can be used by entering a matrix of dissimilarities between objects. Such dissimilarities can be obtained in several ways. Often they are computed from variables that are not necessarily on an interval scale but which may also be binary, ordinal, or nominal. It also happens that dissimilarities are given directly, without resorting to any measurement values [97].

The algorithm consists of two phases: the *build* phase in which an initial set of k representative medoids is selected and the *swap* phase in which a search is carried out to improve the choice of medoids (and hence the cluster allocations). The *build* phase begins by identifying the first medoid, the point for which the sum of dissimilarities to all other

points is as small as possible. Further medoids are selected iteratively through a process in which the remaining points are searched to find that which decreases the objective function as much as possible. Once k medoids have been selected, the *swap* phase commences in which the medoids are considered iteratively. Possible swaps between each medoid and other (non-medoid) points are considered one by one, searching for the largest possible improvement in the objective function. This continues until no further improvement in the objective function can be found. The algorithm is described in detail in [97], pp.102–104. The number of clusters is an explicit input parameter to the PAM algorithm.

Adaptive resonance theory

The adaptive resonance theory (ART) algorithm has three main steps [26]. First, the data are normalised to a unit hypersphere, thus representing only the ratios between the various dimensions of the data. Second, data allocated to each cluster are required to be within a fixed maximum solid angle of the group mean, controlled by a so-called 'vigilance parameter' ρ, namely $X_k \cdot P^i \leq \rho$. However, even when the observation profile and a prototype are closer than the maximum aperture for the group, a further test is applied to ensure that the profile and prototype have the same dominant covariates. This is done in a third step by specifying the extent to which the nearest permissible prototype allocation for the given observation must be on the same side of the data space from the diagonal comprising a vector of ones, $\hat{1}$, using a pre-set parameter, λ:

$$X_k \cdot P^i \leq \lambda X_k \cdot \hat{1}.$$

The ART algorithm is initialised with no prototypes and creates them during each successive pass over the data set. It has some, limited, sensitivity to the order in which the data are presented and converges in a few iterations.

Fuzzy c-means

The fuzzy c-means (FCM) algorithm is a generalisation of the K-means algorithm which is based on the idea of permitting each object to be a member of *every* cluster to a certain degree, rather than an object having to belong to only one cluster at any one time. It is based upon the concept of fuzzy logic promulgated by Zadeh [195] and aims to minimise the objective function

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{i,j})^m \|x_i - v_j\|^2$$

where n is the number of data points, x_i and v_j are the data points and cluster centres and $\mu_{i,j}$ is the membership degree of data x_i to the cluster centre v_j ($\mu_{i,j} \in [0,1]$). m is called the 'fuzziness index' and the value of $m = 2.0$ is usually chosen. An exhaustive description of this method can be found in [11].

For this analysis, the same initialisation technique as used for the K-means algorithm was adopted when considering the Fuzzy c-means and the maximum number of iterations was again set to 100. The Euclidean distance was chosen as metric and the fuziness index was set equal to 2. As for K-means, the number of clusters is an explicit input parameter to FCM.

4.3.2 Cluster validity

Clustering validity is a concept that is used to evaluate the quality of clustering results. If the number of clusters is not known prior to commencing an algorithm, a cluster validity index may be used to determine the best number of clusters for the given data set. Although there are many variations of validity indices, they are all either based on considering the data dispersion in a cluster and between clusters, or considering the scatter matrix of the data points and the one of the clusters centres.

In this study, the following indices were applied to those algorithms for which the number of clusters is an explicit parameter, over a range of number of clusters:

1. Calinski and Harabasz [20]

2. Hartigan [77]

3. Scott and Symons [154]

4. Marriot [118]

5. TraceW [46, 61]

6. Trace$W^{-1}B$ [61]

For each index, the number of clusters to be considered was chosen according to the rule reported in Table 4.3 where i_n is the validity index value obtained for n clusters [183].

Index	Decision rule
Calinski and Harabasz	$\min_n((i_{n+1}-i_n)-(i_n-i_{n-1}))$
Hartigan	$\min_n((i_{n+1}-i_n)-(i_n-i_{n-1}))$
Scott and Symons	$\max_n(i_n-i_{n-1})$
Marriot	$\max_n((i_{n+1}-i_n)-(i_n-i_{n-1}))$
TraceW	$\max_n((i_{n+1}-i_n)-(i_n-i_{n-1}))$
Trace$W^{-1}B$	$\max_n(i_n-i_{n-1})$

Table 4.3: Different validity indices and their associated decision rules

4.3.3 Derivation of classes

Concordance among solutions was evaluated using the Cohen's kappa coefficient κ [30]. This coefficient is a statistical measure of inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance.

To enable visualisation, the original data space (consisting of a large number of dimensions) is transformed by principal component analysis (PCA) [86], and then the points are plotted at their projected position on axes of the first and second principal components. As PCA transforms data such that the first principal component (PC) carries the maximum amount of variance in the data and the second PC carries the next largest variance (etc.),

such a plot, called 'biplot', provides a picture in which the clusters have been 'spread out' as much as possible.

The previously obtained clustering results from Abd El-Rehim and colleagues [1], the cluster validity indices (where appropriate), visualisation of the new clustering results themselves, and the concordance among clustering solutions were then all used heuristically to guide the formulation of a set of rules to define core class membership from the various cluster assignments.

4.3.4 Characterisation of classes

Class characterisation by visualisation

For inspection of the patient characteristics in each class, the distribution of each variable in the class is compared with its distribution in the total sample, using boxplots. A boxplot shows the median expression level (solid horizontal bar), the upper quartile and lower quartile range (shaded grey bar), the highest non-outlier and lowest non-outlier (smaller ticks joined by dashed lines), and any outliers (open circles). For a full description of boxplots, including the statistical definition of outliers see, for example, [171].

Class characterisation by OSRE (orthogonal search rule extraction)

Orthogonal Search Rule Extraction (OSRE) [52] is a computationally efficient algorithm to search for hypercubes in data space, since they map directly onto Boolean rules. This is achieved by modelling the cluster allocation index using an analytical statistical classifier to best fit the cluster membership indicator label, followed by a structured search for the directions in which each data point can be moved before it hits the fitted decision surface for cluster membership. This methodology initially returns a rule for each data point, which triggers a pruning process to keep only those rules which represent large proportions of the data in the clusters, i.e. have high sensitivity for cluster membership, and do so with minimal mixing between clusters, i.e. also have high specificity. The result is a set of multivariate rules involving relatively few covariates, that is to say, low-order rules

containing the covariates that characterise the sub-group of the cluster. The proposed interpretation is that these rules identify the drivers for cluster allocation, which may vary across the cluster but are, in general, well-defined [52].

Note that this method contrasts with widely used rule induction methods in two ways: firstly, there are no univariate cut-offs for groups of data, as in OSRE a sequential univariate search is carried out at the level of each individual data point which returns a multivariate hyperbox around that point, without the need to partition the data along a sequence of univariate covariates; and secondly, that the rules are overlapping, rather than constrained to mutual exclusivity as is usually the case in rule tree induction. Mutually exclusive trees can be readily derived from overlapping rule sets by sequential conjunctions of each rule and the complement of the previous rule along the tree branch, but this loses the benefit of the simplicity of interpretation that comes with the derivation of low-order rules.

Class characterisation by ANN (Artificial Neural Networks)

A conventional multi-layer perceptron artificial neural network (MLP-ANN) model was utilised such that individual H-scores (see Section 3.4.1) derived from the tissue microarray analysis of the clinical samples were set as inputs and the class was set as the output using Boolean notation (i.e. 1 represented membership of a given class, 0 represented non-membership). This allowed the identification of markers that drive membership of a given class and that discriminate the class from the others. A three-layer MLP-ANN (featuring eight nodes in the hidden layer) with a back-propagation algorithm and a sigmoid activation function was used. Learning rate and momentum were set at 0.1 and 0.5 respectively. The approach used in this work is similar to the ones used in [125] and [119].

4.4 Results

4.4.1 Clustering results

HCA, K-means, PAM and ART

The HCA results from Abd El-Rehim *et al.* [1] were utilised, unaltered. Both the K-means and PAM algorithms were run with the number of clusters varying from two to twenty, as the number of clusters is an explicit input parameter of the algorithms. Given that the K-means algorithm can be sensitive to cluster initialisation and in order to obtain reproducible results, this technique was initialised with the cluster assignments obtained by hierarchical clustering. For the ART algorithm, the parameters were set in order to obtain six clusters in order to match the number of clusters previously obtained by HCA. The best validity index obtained for repeated runs of the algorithm with 20 random initialisations was used to select the final clustering assignment.

Fuzzy C-means

Two different (independent) implementations of the fuzzy c-means algorithm were run on the data set in an attempt to obtain clusters, but the algorithm did not perform as hoped. When the number of clusters was set as two and three, it appeared that reasonable results were obtained. However, from examination of the membership function of each point assigned to these clusters, it could be seen that it was very close to either ½ or ⅓, respectively. In other words, every data point was assigned to all the clusters with the same membership. Moreover, when the number of clusters was above three, non-zero memberships were evident for only three clusters and these memberships were similar to the three cluster solution – i.e. for $n > 3$, the $n = 3$ cluster solution was obtained, but with $n - 3$ empty clusters. These results indicated that the fuzzy c-means was not able to obtain clear cluster partitions. The fuzziness index m was altered in an attempt to improve the results obtained, but it was found that little difference in the results was observed until m was close to one. Given that, when $m = 1$, fuzzy c-means is equivalent to

K-means, this result was not useful. As there are many applications for which the fuzzy c-means technique has been successful (see, for example, [176]), these results are not easy to explain, but they may have been caused by the fact that the data under investigation contain a lot of values close to the extremes of each variable. Although the fuzzy c-means algorithm is widely used in literature, it was dropped from further analysis due to its poor performance on the Abd El-Rehim *et al.* dataset.

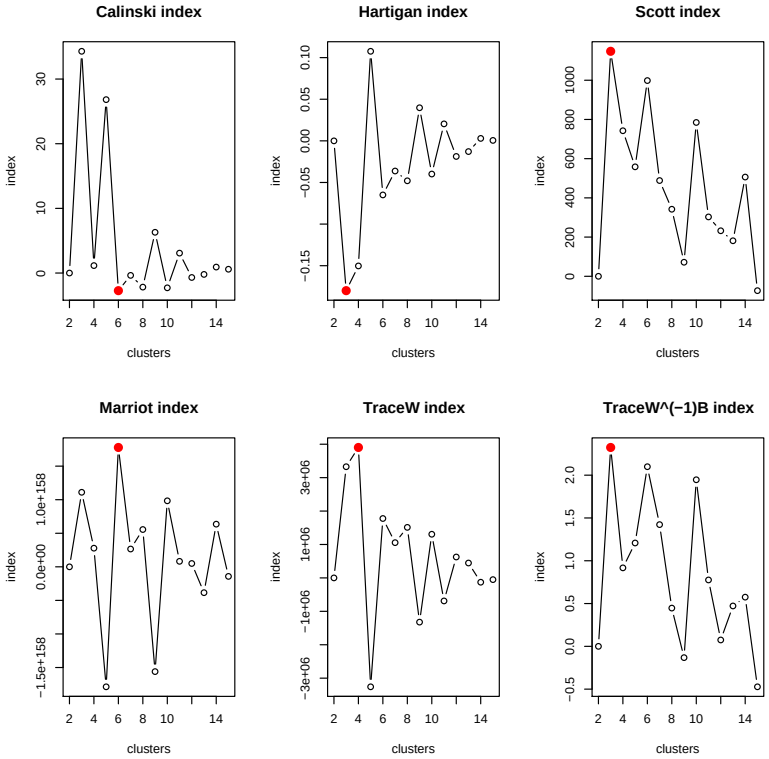
4.4.2 Cluster validity

The values of the decision rule obtained for various values of the validity indices for both K-means and PAM, for 2 to 20 clusters are shown in Figure 4.1; (a) shows the validity decision rule values obtained for K-means and (b) shows those obtained for PAM. The best number of clusters according to each validity index, for each clustering algorithm, is shown in Table 4.4 This corresponds to the either the maximum or minimum decision rule value (depending on the index), as indicated by the red point in Figure 4.1.

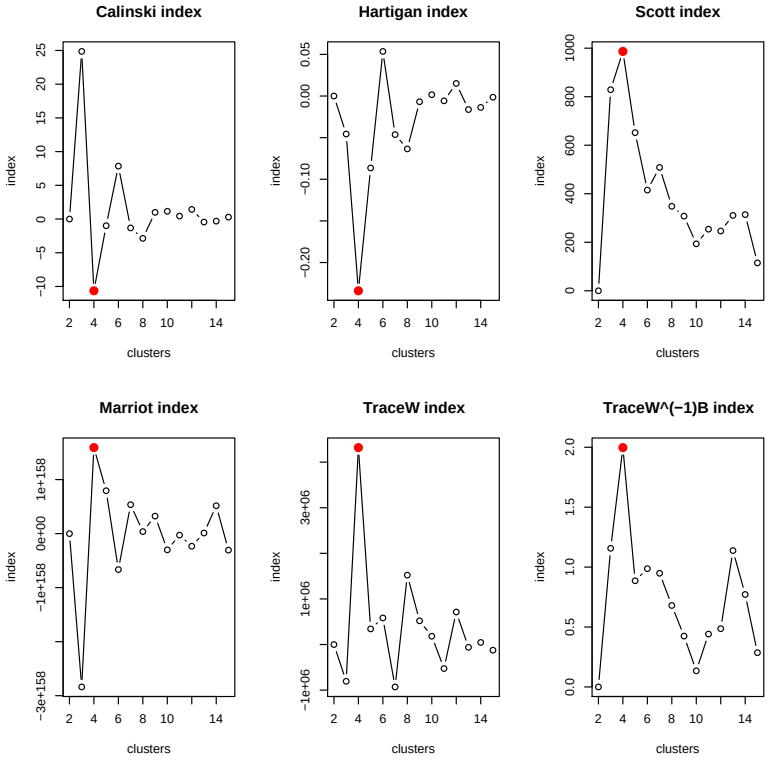
Index	K-means	PAM
Calinski and Harabasz	6	4
Hartigan	3	4
Scott and Symons	3	4
Marriot	6	4
TraceW	4	4
$\text{TraceW}^{-1}\text{B}$	3	4
Minimum sum of ranks	6	4

Table 4.4: Optimum number of clusters estimated by each index for K-means and PAM methods

It can be seen that, while there was not absolute agreement among the indices as to which was the best number of clusters for the K-means method, there is good agreement that the best number of clusters for the PAM method is four. Although the best number of clusters varies according to validity index for K-means, on further inspection, it can be seen from Figure 4.1 that there is more agreement than might be immediately apparent. For example, the Scott and Symons index (which indicated that the best number of clusters was three) indicated that the second best number of clusters was six. Consequently, the



(a) K-means indices behaviors



(b) PAM indices behaviors

Figure 4.1: Cluster validity indices obtained for K-means and PAM clustering, for varying cluster numbers from 2 to 20

	K-means	ART	PAM
HCA	0.497	0.296	0.325
K-means	—	0.494	0.420
ART	—	—	0.224

Table 4.5: Kappa index among different classification

	K-means	ART	PAM
HCA	0.548	0.401	0.332
K-means	—	0.599	0.525
ART	—	—	0.376

Table 4.6: Weighted kappa index among different classification

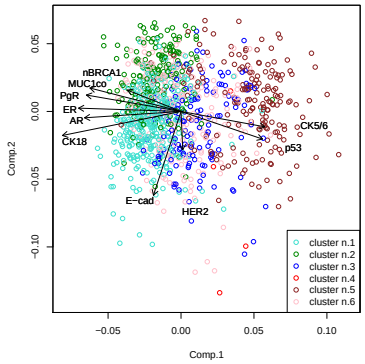
indices were used to rank order the number of clusters and the minimum sum of ranks was examined. It was found that the minimum sum of ranks (a form of consensus among the indices) indicated that the overall best number of clusters was six for K-means and four for PAM. Furthermore, careful examination of Figure 4.1(b) confirms that the six cluster solution for PAM is of relatively poor quality.

4.4.3 Derivation of classes

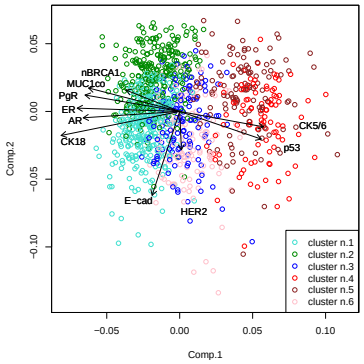
The correspondence of patients assigned in the six cluster solution for each of the methods was then examined. Cohen's kappa and weighted-kappa indices were computed to measure the degree of agreement among algorithms. For the weighted-kappa index, weights were set in decreasing order from one (perfect agreement) to zero (complete disagreement) with a 0.2 step between levels. Results are reported in Tables 4.5 and 4.6. From these tables, a better agreement between K-means and hierarchical algorithms is evident compared to that between ART and hierarchical. It is also evident that the PAM six cluster solution has lower concordance with the original HCA results than either K-means or ART, and that the concordance of PAM with K-means and ART is also correspondingly lower.

The cluster numbers were aligned with those obtained previously by Abd El-Rehim *et al.* in [1] in order to minimise differences and to aid visualisation. Biplots of the aligned clusters are shown in Figure 4.2 for the six cluster solution from each algorithm.

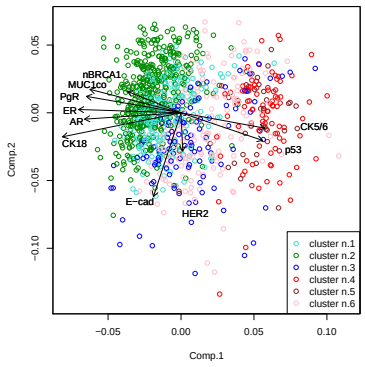
From these plots, it can be seen that the most similar results were obtained from the Hierarchical, K-means and ART. In fact, all these three methods obtain two clusters (1 & 2) split over the left-hand side of the biplots. A third cluster (cluster 6) is evident towards the bottom of the biplot. Then various splits of remaining data into three clusters (3, 4 & 5) can be seen. The PAM algorithm, instead, obtains three clusters (1, 2 & 4) split over the left-hand side, one group is visible towards the bottom (cluster 6) and one is spread in the centre of the biplot (cluster 3). PAM places all patients on the right-hand side into a single cluster (cluster 5).



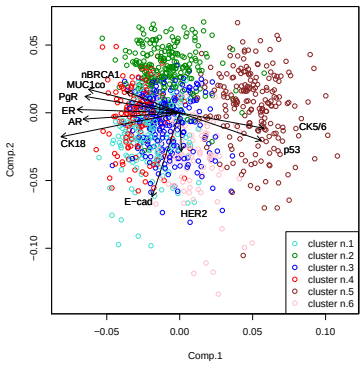
(a) Hierarchical clustering



(b) K-means clustering



(c) ART clustering



(d) PAM clustering

Figure 4.2: Biplots of clusters projected on the first and second principal component axes

The variables under investigation are displayed using vectors (arrows) and their lengths in the plots are related to the expression of the biomarkers. In addition, clusters may be 'informally' characterised by looking at those arrows which overlap a particular set of

points. For example from Figure 4.2(d), the fifth cluster, represented by brown circles, can be characterised by the over-expression of the CK5/6 and p53 markers.

The biplots further confirm that the six cluster solution obtained from the PAM algorithm was the most dissimilar among the considered techniques. Taking into account the results of validity indices analysis, the concordance analysis and the visual analysis, it was decided to remove the six clusters determined by PAM from further analysis.

The cluster distributions (number of patients in each cluster) obtained for the original hierarchical clustering and those obtained for the K-means and ART methods are shown in Table 4.7.

Cluster	HCA	K-means	ART
1	336	301	238
2	180	282	408
3	139	138	111
4	4	97	96
5	183	124	35
6	234	134	188

Table 4.7: Number of cases in each cluster

Focusing on these cluster correspondences, the aim was to define core classes containing the biggest possible number of patients. In a first attempt, considering agreement among the three clustering techniques (HCA, KM and ART) and looking at those patients assigned to the same group by different methods, a total of 382 patients were classified if hierarchical group 4 was considered and 463 if not. After that, for each labelled group, concordances between all pairs of methods were analysed. It was found that the sum of the number of patients assigned to the same group ranged between 459 (pairing HCA and ART) and 645 (pairing KM and ART). These results are again reflected in Table 4.5. Two principles were used to guide the definition of consensus classes: (i) to consider all the clustering techniques analysed and (ii) to get the highest number of patients assigned to any class. These principles conflict, in that strict application of the first principle leads to a decrease in the number of patients assigned to classes. Hence, a heuristic trade-off between the two was employed. As a result, hierarchical group 4 was omitted (being

replaced by group 5), and the ART assignments were not considered in a strictly conjunctive manner. Consequently, a set of six core breast tumour classes was derived following the specific rules reported in Table 4.8, in which the resultant number of patients in each class is shown.

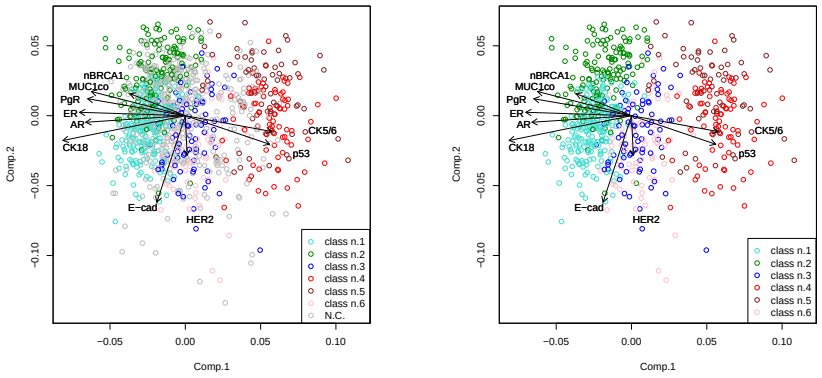
It was found that almost the 62% of data was assigned to these core classes; the remaining patients were placed into a 'not classified' (NC) group. It must be stressed that the derivation of class assignments was made on the basis of the clustering results alone (which are, obviously, based on the 25 markers only) – class assignments, although somewhat subjective, were made *blind* to all clinical and outcome data. It should be also noted that around a third (actually 38%) of all patients were not assigned to any of the core classes.

If cluster ...	Class	No. of cases
H1 & KM1 & (ART1 | ART2)	1	202
H2 & KM2 & (ART1 | ART2)	2	153
H3 & KM3	3	80
H5 & KM4 & ART4	4	82
H5 & KM5	5	69
H6 & KM6 & ART6	6	77
Total number of cases assigned to classes 1–6		663
Total number of cases not classified		413

Table 4.8: Rules for determining consensus classes

4.4.4 Characterisation of classes

Biplots of the six consensus classes were produced and are reported in Figure 4.3, which provides a visualisation of the separation of the classes. Figure 4.3(a) shows the biplot obtained for all patients, in which the cases not assigned to any class (NC) have been coloured grey. It can be seen that these fall mainly into the centre region of the biplot. Figure 4.3(b) shows the biplot obtained for only patients assigned to classes 1 – 6. It can be seen that the classes appear more spread out. The first axis was mainly determined, on the left, by luminal markers including luminal cytokeratins (CK18, CK7/8, CK19), hormone receptors (ER, AR, PgR), and MUC1 over-expression and, on the right, by basal



(a) For all patients (b) For only patients in classes 1–6

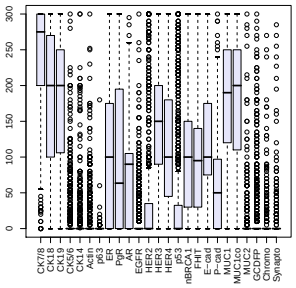
Figure 4.3: Biplots of classes projected on the first and second principal component axes

cytokeratins (CK14 and CK5/6) and partly by p53 over-expression. The second axis is determined, on the top, partly by nuclear BRCA1 (nBRCA1) over-expression and, on the bottom, by HER2 and E-cad over-expression (also HER3 and HER4, although these are not shown as they overlap HER2).

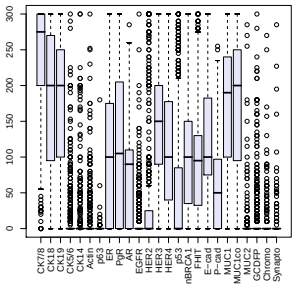
Figure 4.4 shows boxplots of all 25 markers, (a) for all cases, (b) for those cases assigned to classes 1–6, and (c-h) for each class separately. By inspection of both the biplots and the boxplots, a description of each class could be derived. For example, classes 1 and 2 are characterised by strong expression of the luminal CK markers, as well as moderate to strong MUC1 expression (as per the population). However, there is a distinct difference regarding HER3 and HER4 expression. It can also be seen that, classes 4 and 5 both exhibit higher expressions of the basal CKs (CK5/6 and CK14). Triple negative patients with high p53 levels are grouped in class 4, whereas class 5 consist of triple negative patients with low p53 levels. A summary of the class characteristics obtained by visual inspection of the boxplots is given in Table 4.9.

The results obtained from the automated characterisation methods (MLP-ANN and OSRE) are reported in Table 4.10.

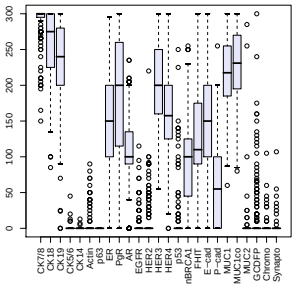
A proposed summary of the essential characterisations of the classes obtained is given in Figure 4.5, according to the available bio-pathological knowledge. It is worth noting that class 2, labelled as Luminal-N, and the split of the basal group into two different sub-



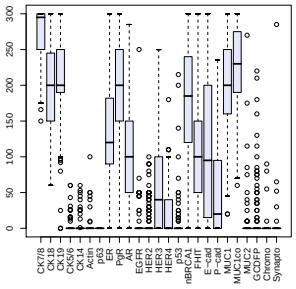
(a) For all patients



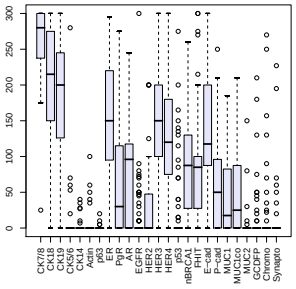
(b) For patients in classes 1–6



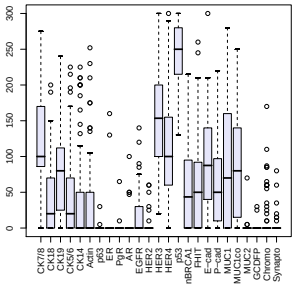
(c) Class 1



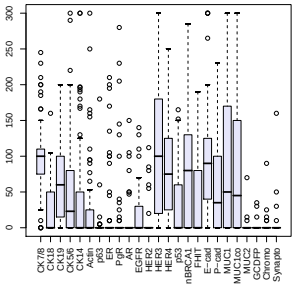
(d) Class 2



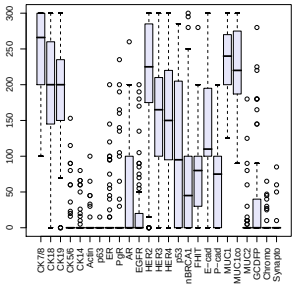
(e) Class 3



(f) Class 4



(g) Class 5



(h) Class 6

Figure 4.4: Boxplot for all markers, whole data and grouped by class

Class	Over-expressed	Under-expressed	Other
1	ER, AR, PgR, HER3, HER4		
2	ER, AR, PgR, nBRCA1	HER3, HER4	
3	ER, AR	MUC1, MUC1co	PgR normal
4	p53	ER, PgR, HER2, MUC1, MUC1co, CK18, CK7/8, CK19	
5		ER, PgR, HER2, MUC1, MUC1co, CK18, CK7/8, CK19	p53 absent
6	HER2, HER3, HER4		ER, AR, PgR absent; p53 widely spread

Table 4.9: Description of classes as determined by statistical characterisation

Class	*Over-expressed*	*Under-expressed*
1 (ANN)	PgR, HER3, HER4, MUC1co	
1 (OSRE)	PgR, HER3, HER4, CK18, CK19, MUC1co	HER2
2 (ANN)	PgR, nBRCA1	HER3, HER4
2 (OSRE)	PgR, nBRCA1, MUC1co	HER3, HER4
3 (ANN)	ER	MUC1
3 (OSRE)	CK7/8, CK18	
4 (ANN)	p53	
4 (OSRE)	HER3, p53	ER, HER2
5 (ANN)	CK5/6	CK7/8
5 (OSRE)		p53; CK7/8, CK19 or HER2, HER4
6 (ANN)	HER2	
6 (OSRE)	HER2, p53, MUC1co	ER

Table 4.10: A summary of rules obtained from the automated methods for defining class memberships

groups depending on p53 levels, appear to be novel findings which were not emphasised in literature yet.

4.5 Clinical Evaluation

4.5.1 Patient clinical outcome

Follow-up data was available for 974 patients, with overall survival ranging from 4 to 224 months (median 123 months, mean 118 months). During this period, a total of 346 patients died, 263 from breast cancer. Patient age ranged from 18 to 72 years (median 54 years). Of the available cases, 708 (66%) cases were aged 50 years or more. At the time of diagnosis, 160 (14.9%) tumours were histological grade 1, 343 (31.9%) grade 2

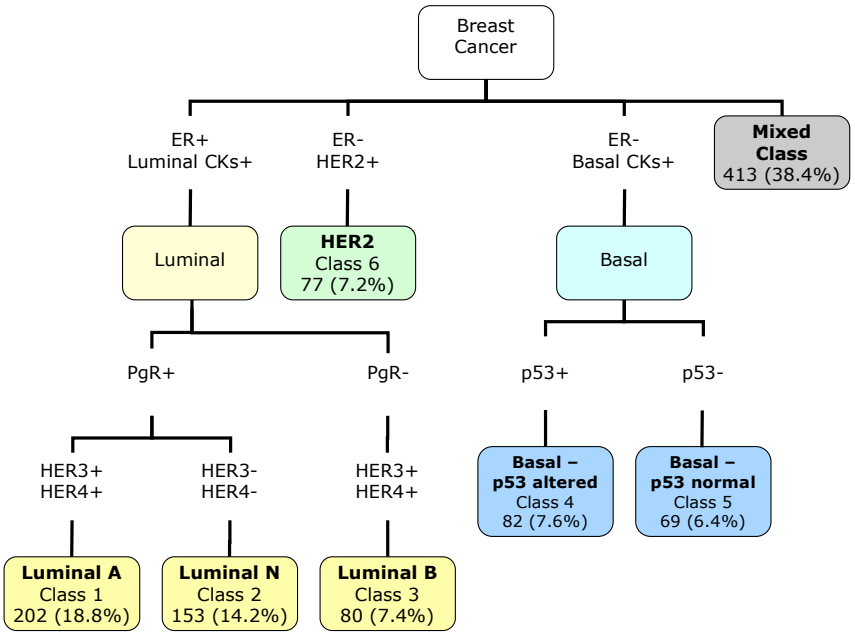


Figure 4.5: A summary of the classes of breast cancer obtained, with indicative class interpretations

and 572 (53.2%) grade 3. A total of 654 (60.8%) patients had lymph node-negative disease and 419 (38.9%) had positive lymph nodes (332 cases with between one and three positive nodes, 87 cases with four or more positive). Frequencies for histological tumour types were: 649 invasive ductal carcinomas of no special type (NST), 171 tubular mixed carcinomas, 30 medullary carcinomas, 112 lobular carcinomas, 27 tubular carcinomas, 11 mucinous carcinomas, five cribriform carcinomas, three papillary carcinomas, 37 mixed NST and lobular carcinomas, 24 mixed NST and special type carcinomas and four miscellaneous tumours. A total of 736 (68.4%) had tumour size more than 1.5 cm and distant metastases was observed in 111 cases.

4.5.2 Clinical characterisation of patients by class

Significant associations, as expected, were found between the classes with respect to patient age, tumour grade, size, lymph node stage and histological tumour type (see Table 4.11). As for grade, there was a relevant distribution of tumours among the six different phenotypic classes: the vast majority of grade 1 tumours were in either Class

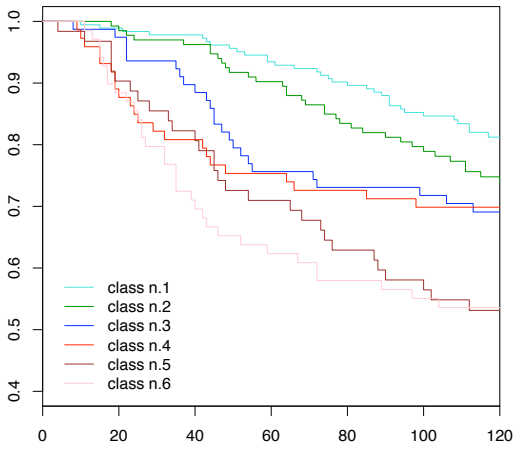


Figure 4.6: Kaplan-Meier curves for ten years survival by class

1 or Class 2. In contrast, the majority of tumours in Classes 3-6 were of higher grade (grade 2 or 3). Tumours with high values of grade are known to be more aggressive and to have a shorter survival time [48]. Therefore, Classes 1 and 2 appeared to express good prognostic factors.

Figure 4.6 shows ten year breast cancer specific survival for the six biological classes. It can be seen that the different classes were distinct with respect to overall survival. The highest frequency of breast cancer related mortality was seen in patients whose tumours belonged to classes 5 (*basal-p53-normal*) and 6 (HER2). A lower, but still high, frequency was seen in patients with tumours from both classes 3 (luminal B) and 4 (*basal-p53-altered*). Classes 1 (luminal A) and 2 (luminal N) had the lowest frequency of death due to breast cancer.

A boxplot of the Nottingham Prognostic Index (NPI) split by class is shown in Figure 4.7. It can be seen that the NPI for classes 1 and 2 is lower than that of classes 3–6 (overall Kruskal-Wallis $p << 0.001$). It can also be seen that classes 1 and 2 have similar NPI, and classes 3–6 have similar NPI (to each other). This is an interesting observation for two reasons. Firstly, it confirms that the NPI is providing discriminant information between classes 1 and 2, and classes 3–6. Secondly, it suggests that the class divisions are providing *additional* information to the NPI.

	Breast Cancer Class						
	1	2	3	4	5	6	ϕ
Age							
≤ 50	76 (37.6)	63 (41.2)	24 (30.0)	55 (67.1)	33 (47.8)	37 (48.1)	0.209
> 50	126 (62.4)	90 (58.8)	56 (70.0)	27 (32.9)	36 (52.2)	40 (51.9)	
Total	202	153	80	82	69	77	
Grade							
1	58 (28.9)	43 (28.1)	2 (2.5)	0 (0)	2 (2.9)	1 (1.3)	
2	81 (40.2)	89 (58.2)	18 (22.5)	1 (1.2)	7 (10.1)	12 (15.6)	0.660
3	62 (30.8)	21 (13.7)	60 (75.0)	81 (98.8)	60 (87.0)	64 (83.1)	
Total	201	153	80	82	69	77	
Size							
$\leq$ 1.5cm	79 (39.1)	65 (42.5)	20 (25.0)	12 (14.6)	15 (21.7)	16 (20.8)	0.225
$>$ 1.5cm	123 (60.9)	88 (57.5)	60 (75.0)	70 (85.4)	54 (78.3)	61 (79.2)	
Total	202	153	80	82	69	77	
Lymph Node Stage							
1	132 (65.3)	108 (70.6)	39 (48.7)	50 (61.0)	52 (75.4)	36 (46.8)	
2	58 (28.7)	37 (24.2)	35 (43.8)	23 (28.0)	10 (14.5)	30 (39.0)	0.217
3	12 (5.9)	7 (4.6)	6 (7.5)	9 (11.0)	7 (10.1)	10 (13.0)	
Total	202	152	80	82	69	76	
Tumour type							
Invasive ductal/NST	97 (48.0)	45 (29.4)	64 (80.0)	70 (85.4)	53 (76.8)	68 (88.3)	
Tubular mixed	52 (25.7)	50 (32.6)	8 (10.0)	0 (0)	1 (1.5)	5 (6.5)	
Medullary	0 (0)	0 (0)	0 (0)	10 (12.2)	5 (7.2)	2 (2.6)	
Lobular	18 (8.9)	34 (22.2)	6 (7.5)	0 (0)	4 (5.8)	1 (1.3)	0.622
Special types	19 (9.4)	11 (7.2)	0 (0)	1 (1.2)	0 (0)	0 (0)	
Mixed NST & lobular	6 (3.0)	7 (4.6)	2 (2.5)	1 (1.2)	3 (4.3)	0 (0)	
Mixed NST & special type	9 (4.5)	5 (3.3)	0 (0)	0 (0)	1 (1.5)	0 (0)	
Miscellaneous	0 (0)	1 (0.7)	0 (0)	0 (0)	2 (2.9)	0 (0)	
Total	201	153	80	82	69	76	

Table 4.11: Breast Cancer Class distribution in relation to clinicopathological parameters (NST: No Special Type)

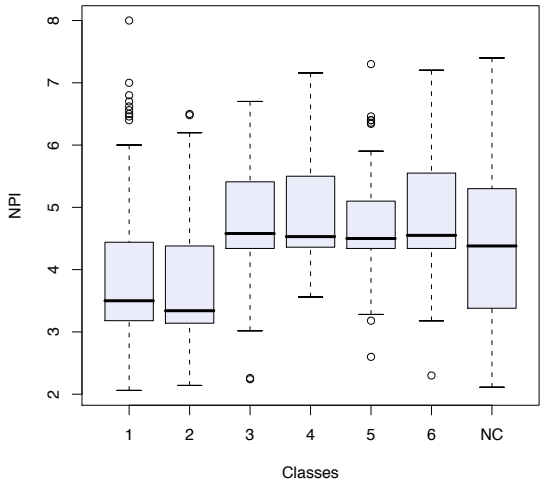


Figure 4.7: Boxplots of Nottingham Prognostic Index (NPI) by class

4.5.3 Comparison between the six classes and the ones identified in previous studies

The six classes identified in this work can not be used to define breast cancer phenotypes, as a consistent number of patients presented mixed class characteristics. However, it was thought that a comparison with classes previously identified in similar works would be interesting and could help for a clearer interpretation of the results. In particular, the six classes identified in this study were compared with those defined in the previous work of Abd El-Rehim *et al.* [1], and with the ones identified in Ambrogi *et al.* [5] and Sørlie *et al.* [158]. The groups identified by consensus clustering included three luminal tumour classes (classes 1, 2 and 3) characterised by high luminal cytokeratin expression and the expression of hormone receptors. There were two basal tumour classes (class 4 and 5), characterised by low luminal cytokeratin expression and a triple negative phenotype (i.e. ER, PgR and HER2 negative). Lastly, there was the HER2 class (class 6) characterised by high luminal cytokeratin and HER2 expression. These classes are similar to those determined by gene expression profiling, but in this study the definitions of the luminal and basal tumours were refined into further distinct classes with different clinical outcome.

Class 1 tumours are consistent with the previously identified group 1 [1] and Ambrogi's group 1 [5], but are more distinctly defined. These tumours have high expression

of luminal cytokeratins (CK7/8, CK18 and CK19), hormone receptors, HER3 and HER4 and show high homology to the luminal-A type tumours, as identified in gene array studies. Class 3 tumours were also characterised by high levels of HER3 and HER4 but, in contrast, showed relatively lower levels of PgR (levels of ER and AR were similar). This class was not identified by the previous study using the same panel of markers [1], but is similar to group 2 in the Ferrara series [5]. This class of tumour shows homology to the luminal-B group of tumours [158, 160].

Whilst gene expression profiling has determined two luminal tumour classes, in this study those tumours with a luminal phenotype have been divided into a further class (class 2). This novel class of tumours, which was designated as the luminal-N class, whilst having high levels of ER and PgR, has negative/low expression of the EGFRs, particularly HER3 and HER4. Interestingly, the class 1 (luminal-A) and class 2 (luminal-N) tumours were similarly associated with good prognostic factors, including smaller tumour size, grade 1 tumours, node-negative and tubular mixed carcinomas. The luminal-N tumours were further distinct from the luminal-A tumours by high expression of nuclear BRCA1 protein. In contrast, the third luminal class, luminal-B, although phenotypically similar to the luminal-A tumours (except for PgR expression), consisted of those tumours with poorer prognostic factors such as larger tumour size, higher stage and grade. It is also apparent that HER3 and HER4 are important discriminators in the breast cancer classification presented here, although there remains controversy as to their prognostic significance [3, 190].

Previously, a basal-like subtype was identified using protein expression (Abd El-Rehim's group 5 [1] and Ambrogi's group 3 [5]). However, it is now thought that the basal-like subtype is heterogeneous [105] and, indeed, two basal-like classes (classes 4 and 5) have now been determined. These classes were characterised by high expression of basal cytokeratins (CK5/6 and CK14), low expression of luminal cytokeratins and a triple negative phenotype. They were, however, separated by p53 protein expression levels: the basal-p53-altered tumours (class 4) expressed high p53 and basal-p53-normal

			Previous publications		
Class	Name	Key Features	Abd El-Rehim *et al.* [1]	Ambrogi *et al.* [5]	Sørlie *et al.* [158]
1	*Luminal-A*	luminal, HER3/4+	Group 1	Cluster 1	luminal A
2	*Luminal-N*	luminal, HER3/4–	Group 2	—	
3	*Luminal-B*	luminal, PgR–	—	Cluster 2	luminal B
4	*Basal-p53-altered*	basal, p53+	Group 5	Cluster 3	basal
5	*Basal-p53-normal*	basal, p53-	Group 5	—	
6	*HER2*	luminal, HER2+	Group 3 & 6	Cluster 4	HER2
NC	—	—	Group 4	—	—

Table 4.12: Comparison of breast cancer classes determined in this study compared to those previously identified

tumours (class 5) had low p53. High frequency of tumour suppressor p53 mutations and protein expression have previously been detected in the basal-like subtype [49, 138, 158]. A higher proportion of younger patients had basal-p53-altered tumours compared with basal-p53-normal tumours and all but one were grade 3. Basal p53 altered tumours primarily consisted of ductal NST and medullary tumours, whilst in basal-p53-normal, there was a wide range of histological types. The association between medullary carcinomas, p53 and basal tumours has been previously demonstrated [33, 158].

Those tumours with high HER2 expression were clustered into one class (class 6), which is homologous to Sørlie's HER2 group [158] and Ambrogi's group 4 [5]. This class has the worst overall survival. A summary of the new clinical phenotypes and their relationship to those previously identified is shown in Table 4.12.

While 25 protein markers were originally used to derive the classes, not all will necessarily be needed in order to adopt these classes in clinical decision making. Having established core classes, it is currently investigating how a new patient may be reliably assigned into a class using the minimum number of markers. As indicated in Table 4.9 and Figure 4.5, only around eight–twelve of the markers appear to be key drivers. The

proportion of patients that would remain 'unclassified' in such a model-based class assignment is not necessarily 38%. It may well be that, for example, all patients could be classified into one of these six classes, but with varying degrees of certainty. That is, 'unclassified' in this study does not necessarily mean 'unclassifiable' in clinical practice. Actually, the core classes appear from the merging of results from different clustering procedures. Further studies are required to validate these classes and to enable the creation of a clinically usable algorithm for prospective classification, taking into account current therapeutic strategies.

4.6 Discussion

This study has extended the previous work [1], with the application of different clustering techniques to address the issue of the non-existence of the 'perfect' clustering algorithm. In particular, in this work four different clustering methods (in addition to the hierarchical method used in [1]) were applied to a multidimensional dataset of protein biomarker data, in order to evaluate the stability of results coming from different techniques. Different clustering algorithms result in different clusters, particularly when large multi-dimensional data sets are considered.

To explore the extent of the differences among different algorithms, an informal consensus clustering was used, grouping together patients that were assigned to 'similar' clusters by different clustering algorithms. The consensus approach was similar to the one used by Kellam *et al.* [98], but instead of building an agreement matrix, the previously published hierarchical clustering solution (and associated labelling) was used as a fixed reference. In this way, a set of six core classes of breast cancer was elucidated. Another important issue that emerges when cluster analysis is performed, is the best number of clusters to consider. Several validity indices were proposed in recent years (see, for example, [183]) to evaluate the compactness of clusters and the separation among them. For the algorithms which take an explicit number of clusters as an input parameter (i.e.

K-means and PAM), cluster validity indices were used to guide the choice of the 'best' number of clusters. Note that cluster validity indices would have been applied to the fuzzy c-means algorithm had it not been dropped from analysis for the reasons outlined in Section 4.4.1.

Furthermore, this study confirmed, as already highlighted in [5], that cluster analysis should be treated with caution, as different clustering algorithms will lead to different groupings of tumours. As reported in [192], most of the proposed clustering algorithms are largely heuristically motivated, but the issue of determining the 'correct' number of clusters and choosing a 'good' clustering algorithm are not yet rigorously solved. In particular, in this study, the PAM algorithm, when run with six clusters as an input, provided groups that were different from those obtained using the other techniques. In addition, the hierarchical algorithm, commonly used in standard bioinformatics applications of cluster analysis, such as [137] or [170], seems to provide a dissimilar and skewed classification with respect to the others, thus reducing the degree of overall concordance and the number of subjects assigned to the core classes.

In conclusion, it has been clearly demonstrated that different clustering algorithms can produce quite different solutions on such multi-dimensional data. A methodology for reaching consensus from the various results that may be obtained from clustering algorithms has been shown, together with the illustration of this consensus methodology on a well-known set of breast cancer data. In doing so, possible new sub-classes of breast cancer which warrant further investigation have been identified. It must be emphasised that this consensus methodology, by its heuristic nature, should be considered as an exploratory technique, and must not be considered as providing any form of definitive answer. Further work exploring, for example, the statistical properties of the considered algorithms may provide relevant information on the structure on this complex biological problem.

4.7 Triple-negative note

Clinically, breast cancer patients fall into three main groups: those with estrogen (ER) and progesterone (PR) receptor-positive tumours, who are cured with anti-hormone treatments with/without chemotherapy; those with HER2 positive tumours, who can receive a HER2-targeted therapy; those with ER, PR and HER2 negative tumours, for whom the lack of tailored therapies makes chemotherapy the only available modality of systemic care [185].

Genome-wide DNA microarray analysis were used to classify breast cancers into five main expression profile groups, two of them ER-positive (luminal A and B) and three ER negative (normal breast-like, ERBB2 [also known as HER2] and basal-like) [137, 158, 159]. Consistently, single nucleotide polymorphism-association studies indicated that different genetic risk factors can be associated to ER-positive or ER-negative tumours, and that they may also vary according to the expression of HER2 or of basal cancer markers [65].

The basal-like cancer group includes tumours that lack both steroid hormone receptors and HER2 expression, the so-called triple-negative cancers [29, 132, 158, 169]. However, despite the clinical similarities between basal-like and triple-negative tumours, including higher incidence in younger patients [36, 166], higher histologic grade [36, 143, 166], aggressive clinical behaviour and poor prognosis [18, 62], triple-negative and basal like breast cancers are not synonymous. Indeed, not all basal-like cancers are negative for ER, PgR and HER2 expression [132] and the triple negative group also encompasses non-basal-like tumours, namely normal breast like cancers [36]. Notably, although normal breast-like tumours have a somewhat better prognosis than basal-like cancers [158, 159, 161], they do not respond to neoadjuvant chemotherapy as well as basal-like cancers do [25, 151].

Salient features of triple-negative breast cancers include overexpression of EGFR and c-KIT, high proliferative rates, frequent genomic alterations, phenotypic similarity to BRCA1-associated cancers and frequent mutations of TP53 [29, 106, 139]. In particu-

lar, p53 appears heterogeneously expressed in triple-negative tumours, suggesting that it may be associated with specific subgroups. Hence, two independent breast tumour case series for p53 expression in triple-negative breast cancers were comparatively analysed.

A series of 633 patients who underwent surgery for primary infiltrating breast cancer between 1983 and 1992 at the University of Ferrara was studied [5]. Immunohistochemistry (IHC)-determined ER, PR, Ki-67/MIB-1 proliferation index (Ki-67), HER2 and p53 levels were analysed. Percent expression values of ER, PR and HER2 tended to distribute around discrete values (0%, 10%, 25%, 50%, 75% and 100% of tumour cells) and were categorized accordingly. Percentages of Ki-67 and p53 expressing cells were analysed without discretization (Table 4.13), but they were reported in categories for convenience (Table 4.14).

Applying non-hierarchical algorithms, Ambrogi *et al.* [5] previously identified four breast tumour clusters distinguishing distinct tumour profiles according to the expression of traditional markers: cluster 1, characterised by high values of ER/PR; cluster 2, with intermediate ER/PR values; cluster 3, with low-to-nil ER and high p53 and cluster 4 with low-to-nil PR and high HER2 values.

In the present analysis, p53 protein expression was shown to be able to subdivide the triple negative Ferrara cases into two distinct subsets, that were tightly associated to clusters 2 and 3, respectively [5]: low to nil p53 levels were only observed in cluster 2 while overexpression of p53 was only seen in cluster 3 (Tables 4.13, 4.14). These findings support the hypothesis that the triple-negative cancers group de facto includes two different biological entities: basal-like (p53-positive) and normal breast-like (p53-negative) tumours. Owing to this dichotomy, p53 expression may critically help identifying the corresponding patient subclusters, and may possess a specific biological/prognostic value.

Looking for additional evidence, the triple-negative tumours of an independent dataset of 1076 patients from the Nottingham Tenovus Primary Breast Carcinoma Series [1], which were evaluated by IHC for 25 markers including p53, were also analysed for basal cancer-related biological markers. Levels of IHC reactivity were categorized using a mod-

Ki 67/MIB-1	P53	Cluster
4.21	0.00	2
75.00	0.00	2
12.10	0.00	2
45.00	2.50	2
72.69	0.00	2
16.11	0.01	2
14.78	8.40	2
4.82	0.00	2
48.98	0.00	2
10.60	3.75	2
18.70	0.00	2
73.38	0.00	2
20.00	0.00	2
1.28	0.00	2
69.61	0.00	2
90.77	0.00	2
54.93	0.00	2
45.39	2.00	2
40.20	75.79	3
14.50	98.00	3
75.00	97.00	3
20.00	81.00	3
14.80	95.00	3
18.23	55.21	3
65.31	98.00	3
45.00	98.00	3
47.60	98.00	3
50.00	98.37	3
75.04	65.00	3
62.62	35.20	3
85.00	95.00	3
42.30	98.00	3
33.93	40.93	3

Table 4.13: Values of Ki-67/MIB-1 and p53 expression levels in triple-negative patients (Ferrara case series). The third column shows the cluster defined in Ambrogi *et al.* [5]

p53	Cluster 2	Cluster 3
0	14	0
1-10	4	0
11-75	0	4
76-100	0	11

Table 4.14: Distribution of p53 expression levels in triple-negative patients

ified Hscore (values between 0-300) that integrates staining intensity and percentages of positive cells. As expected, a decreased expression of luminal phenotype-associated markers (MUC1, CK7/8, CK18, and CK19) and a concomitant increased expression of markers associated with the basal-like phenotype (EGFR, CK5/6, CK14 and p53) were observed (Figure 4.8). Consistently with the Ferrara data, p53 was markedly overexpressed in only a subset of triple-negative cases (Figure 4.9).

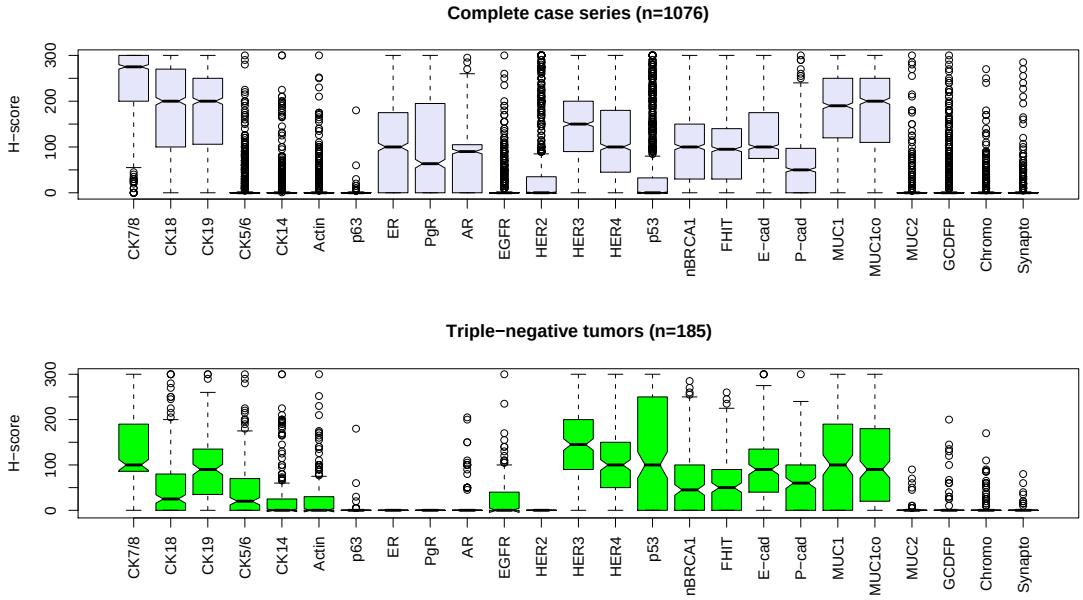


Figure 4.8: Boxplots for Nottingham data

Missense TP53 mutations often lead to higher stability of p53 proteins, that become detectable by IHC [136]. Mutant p53 proteins can become functionally dominant-negative and are characterised by gain-of-function properties. In contrast, truncated p53 proteins are largely unstable, and cannot be revealed by IHC analysis, similarly to wild-type p53. As a consequence, missense TP53 mutations are predominantly IHC positive (92.9%), whereas truncating TP53 mutations are predominantly IHC negative (88.5%) [136]. Breast cancer patients carrying missense TP53 mutations show worse disease-free survival than those with wild-type TP53, whereas women carrying cancers with truncating TP53 mutations do not [136]. Consistently, Langerod *et al.* [108] demonstrated that, among the five breast cancer subgroups identified by Perou *et al.* [137], the basal-like

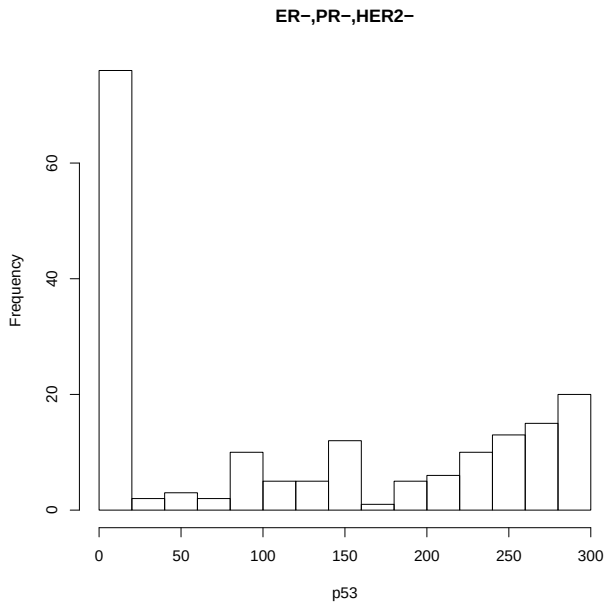


Figure 4.9: Distribution of p53 expression levels in Nottingham case series

and ERBB2+ phenotypes show the highest TP53 mRNA expression while the normal like phenotype has the lowest TP53 mRNA levels. These findings were confirmed at the protein level. Since IHC detection of p53 protein largely identifies missense TP53 mutations, p53 IHC positivity may represent a useful biological marker to discriminate more aggressive triple-negative basal-like tumours from triple-negative tumours with a normal-like phenotype. In addition, as TP53 gene mutations are predictive of response to taxanes in reconstituted model systems [28] and in patients [38, 76, 174], knowledge of p53 status may also provide powerful information to select, among the triple-negative tumours, those more likely to benefit from taxane versus anthracyclines/alkylating agent-based chemotherapy [25, 38, 151]. Taken together, these findings suggest that analysis of p53 expression may help selecting patient subgroups with different biological history among triple-negative breast cancers. This may be associated with powerful predictive information for differential care and clinical trials planning.

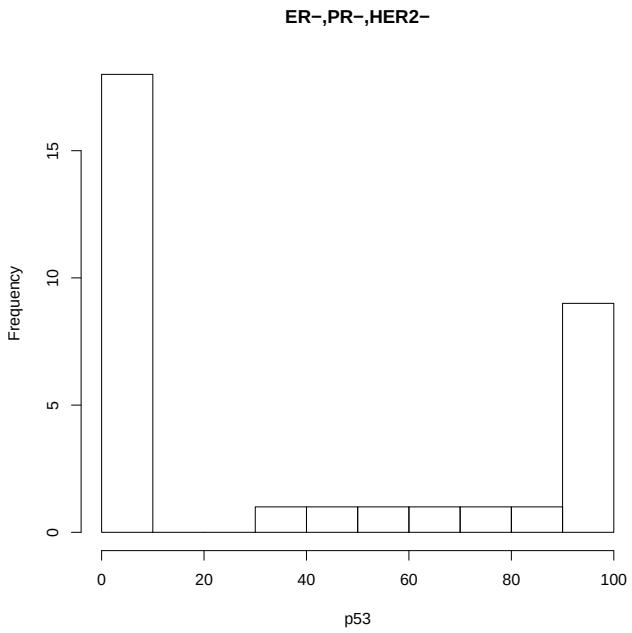


Figure 4.10: Distribution of p53 expression levels in Ferrara case series

4.8 Summary

In this chapter, a comparison between several unsupervised clustering techniques which were run in the attempt to define a set of core breast cancer classes was reported. Through an informal consensus clustering, a set of six groups was identified and several statistical methods have been used to characterise each group. The relationship between several clinical information and the core classes was also studied resulting in some novel findings which have not been emphasised in literature yet. The chapter was concluded by a short note on the role of p53 on triple-negative patients.

In the next chapter supervised classification techniques will be analysed and a novel algorithm, which does not assume any particular distribution for the variables under investigation, will be presented and validated over different data sets taken from the UCI Machine Learning Repository.

Chapter 5

Supervised Classification Techniques

5.1 Background and motivation

The classification of breast cancer patients is of great importance in cancer diagnosis. During the last few years, many algorithms have been proposed for this task and modern machine learning techniques are progressively being used by biologists to obtain proper tumour information from the databases. In this chapter, a review of different supervised machine learning techniques for classification of data sets and a methodological comparison of these are reported.

For the first part of this analysis, the same data on breast cancer [1] used in previous chapter will be considered. The full list of variables is reported in Table 4.2 in Chapter 4. Over this dataset, a C4.5 tree classifier, a Multilayer Perceptron Artificial Neural Network and a naive Bayes classifier will be applied. The same machine learning techniques were already used in literature: in particular, Bellaachia and Guven in [10], revising a study of Delen *et al.* [34], used the above methods to find the most suitable one for predicting survivability rate of breast cancer patients. This study was instead motivated by the necessity to find an automated and robust method to validate the previous classification of breast cancer markers (see Chapter 4). In fact, six classes were obtained using agreement between different clustering algorithms. Starting from these groups, the aim was to re-

produce the classification keeping into account the high non-normality of the data (see Figures 5.1 and 5.2). For this reason, the C4.5 and the Multilayer Perceptron classifiers were used and then the results were compared with the naive Bayes ones. It is important to note that out of the 1076 patients, only 62% (663 cases) were classified into one of the six core groups presented in Chapter 4, while the remaining 38% presented indeterminate or mixed characteristics. In this part of the study, the focus was only on the subset of the 'in-class' cases to run the classifiers on in order to find an automated way to justify and reproduce the classification obtained before. This subset represents a novel clinical categorisation of breast cancer which is interesting in its own right and presents a challenging classification task. Further understanding of undetermined cases is left open for future investigation.

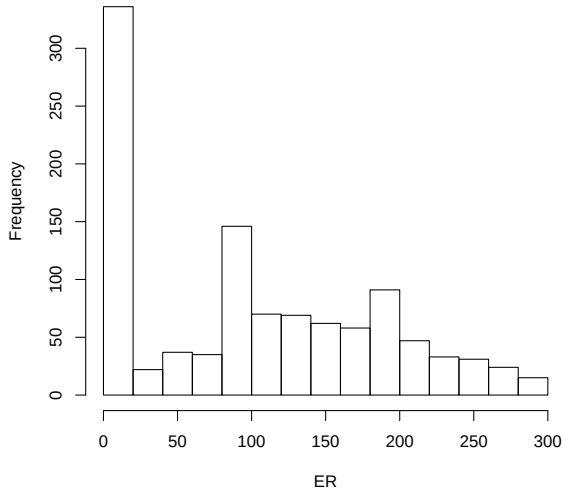


Figure 5.1: Histogram of variable ER

In the second part of the chapter, different case series taken from the UCI Machine Learning Repository [8, 156] will be considered to cope with non normal data. On these data sets, the performance of the naive Bayes classifier will be compared with the Logistic Regression approach for classification. Moreover, as the naive Bayes assumption of normality of the data is strongly violated in many real-world problems, a new method

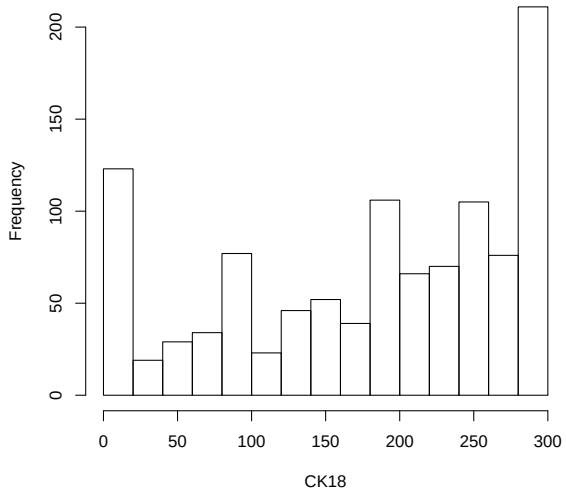
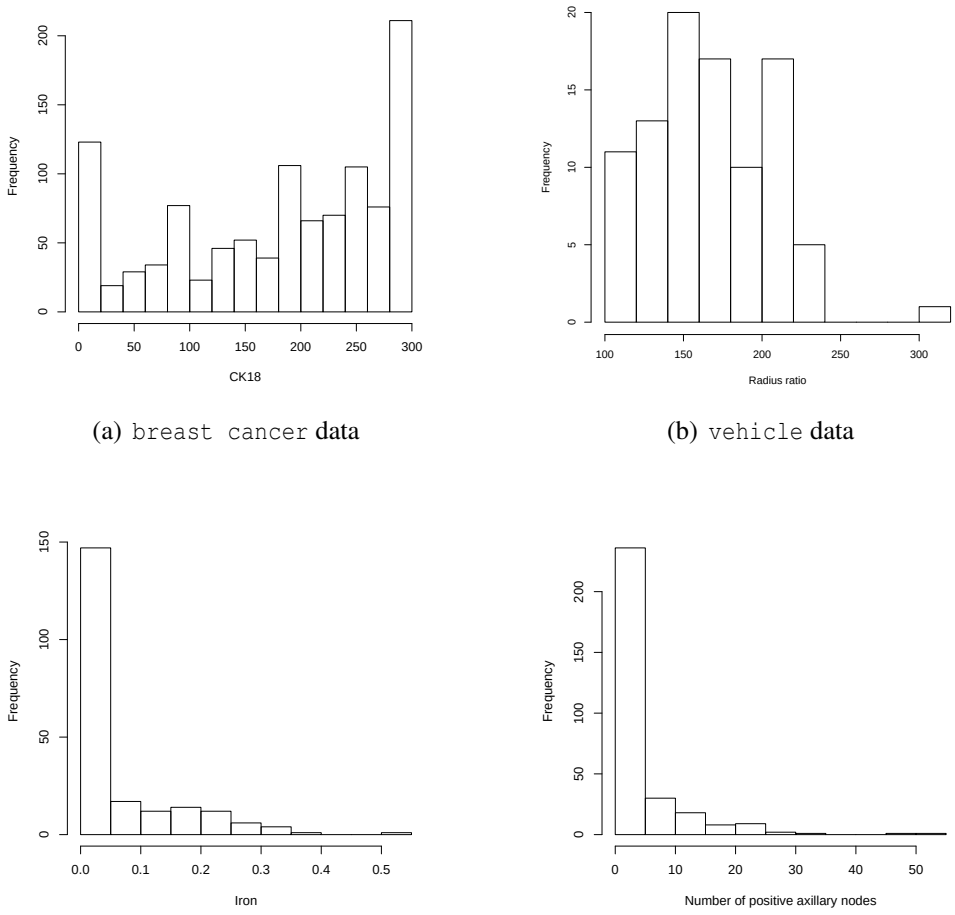


Figure 5.2: Histogram of variable CK18

for the implementation of a Bayesian classifier was developed and will be presented in the last part of this chapter. This method deals with continuous and non-normal variables which, as in the cases presented here, do not follow normal distributions (see Figure 5.3). The algorithm has the same structure as the naive Bayes one, considering the ratio between areas under curves of the variables distribution. The results obtained with the new method will be compared with both those found by the naive Bayes algorithm (variables approximated by a Gaussian distribution) and those obtained by applying a multinomial Logistic Regression model.

The chapter is structured as follows: in Section 5.2, the technical details of the three classifiers considered are reported (a general description of each technique has already been presented in Chapter 2). A short description of the Logistic Regression approach and of the data sets used are also present in Section 5.2. Then in Section 5.3 the results obtained applying classifiers are presented and the differences among them pointed out. Section 5.4 is reserved for the description of the new method developed to cope with the non-normality of the data.



(a) `breast cancer` data

(b) `vehicle` data

(c) `glass` data

(d) `haberman` data

Figure 5.3: Histogram of sample variables

5.2 Experiments settings

The performances of the naive Bayes classifier, the C4.5 decision tree and the Multilayer Perceptron Neural Network were evaluated using the WEKA software [189]. It is a popular suite of machine learning free software written in Java and developed at the University of Waikato in New Zealand. Instead logistic regression and the new method were run using *R* [114]. All the techniques analysed were run 10 times using the 10-fold cross validation option and the accuracy of the obtained classification was evaluated simply by looking at the percentage of the correctly classified instances. The mean of the returning results was then computed.

As C4.5 can handle continuous attributes, there was no need to discretise any of the

attributes and in this work experiments the default values for the parameters were used, with the minimum number of object in each leaf being set as 2. The default version does perform some pruning (using the subtree raising approach), but does not perform error pruning.

For the Multilayer Perceptron classifier the default parameters were again used, leaving the number of neurons in the hidden layer as 15, which is the sum of the number of attributes and classes divided by two. The default backpropagation learning algorithm was used. Comparison of alternative learning algorithms is outside the scope of this work.

The default options were also used when running the naive Bayes algorithm, accepting the option of a normal distribution estimator for numeric attributes.

5.2.1 Logistic Regression

Logistic Regression is an approach to learning functions of the form $f : X \rightarrow C$, or $P(C|X)$ in the case where C is discrete-valued, and $X = \langle X_1 \ldots X_n \rangle$ is any vector containing discrete or continuous variables [129]. Logistic Regression assumes a parametric form for the distribution $P(C|X)$, then directly estimates its parameters from the training data. In this way, the 'two-steps' approach for estimating $P(C|X)$ used by the naive Bayes may be avoided. In this sense, Logistic Regression is often referred to as a *discriminative* classifier, because the distribution $P(C|X)$ can be viewed as directly discriminating the value of the target C for any given instance X. As shown in [129], if C is boolean and the Gaussian Naive Bayes (GNB) assumptions hold, then asymptotically (as the number of training examples grows toward infinity) GNB and Logistic Regression converge toward identical classifiers. However, as demonstrated in detail in [131], GNB parameter estimates converge toward their asymptotic values in order $\log n$ examples, where n is the dimension of X. In contrast, Logistic Regression parameter estimates converge more slowly, requiring order n examples.

When the response variable C is boolean (0 or 1), the Logistic Regression, fitted by a generalised linear model, may be used to model $P(1|X)$ (and then $P(0|X)$ is equal to

$1 - P(1|X)$); a multinomial logistic regression model is instead needed when there are more than two classes.

Measures for predictive accuracy

There are many different measures for assessing the accuracy of a model [75]; two of them are *calibration* and *discrimination*. When a fraction of about P of the events that are predicted with probability P actually occur, it can be said that the predicted probabilities are well calibrated and a suitable model for $P(C|X)$ has been found [172]. Discrimination, instead, measures a predictor's ability to separate patients with different responses [75]. When the outcome variable is dichotomous and predictions are stated as probabilities that an event will occur, calibration and discrimination are more informative than other indices (like, for example, the expected squared error) in measuring accuracy [75]. The calibration plot is a method that shows how well the classifier is calibrated and a perfectly calibrated classifier is represented by a diagonal on the graph [173]. In this work, these plots were produced following the procedure described in [172], plotting the fitted values versus the actual average values.

A *c concordance* index is a widely applicable measure of predictive discrimination and it applies to ordinary continuous outcomes, dichotomous diagnostic outcomes and ordinal outcomes. This index of predictive discrimination is related to a rank correlation between predicted and observed outcomes. The c index is defined as the proportion of all patient pairs in which the predictions and outcomes are concordant. For predicting binary outcomes, c is identical to the area under a receiver operating characteristic (ROC) curve [75].

A ROC curve is a tool to measure the quality of a binary classifier independently from the variation in time of the ratio between positive and negative events [173]. In other words, it is a graphical plot of the *sensitivity* versus *(1 - specificity)* for a binary classifier system as its discrimination threshold is varied. The ROC can also be represented equivalently by plotting the fraction of true positives (TPR = true positive rate) versus the

fraction of false positives (FPR = false positive rate). A completely random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners. Usually, one is interested in the area under the ROC curve, which gives the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. A random classifier has an area of 0.5, while an ideal one has an area of 1.

5.3 Results

As stated previously, in this first part of the study, the classifiers were applied in order to get an automated way to justify and reproduce the classification previously obtained. The results obtained from the C4.5 were quite good, precisely 582 cases were correctly classified (87.8%) and just 81 (12.2%) incorrectly classified. The main concern in using this classifier came from the set of rules that were produced: they appear to be quite numerous and not straightforward, especially if they should be used by scientists not familiar with computational analysis.

The Multilayer Perceptron (MLP) neural network was then considered. This method performed better than the C4.5 succeeding in correctly classifying 647 instances (97.6%) out of 663; just 16 cases (2.4%) were misclassified.

Finally, the naive Bayes (NB) classifier was applied. This method performed worse than the previous ones, classifying properly a smaller amount of cases (576, corresponding to 86.9%). A summary of the above results can be found in Table 5.1.

Whole data

Method	*Classified*	*Misclassified*
C4.5	582 (87.8%)	81 (12.2%)
MLP	647 (97.6%)	16 (2.4%)
NB	576 (86.9%)	87 (13.1%)

Table 5.1: Comparison of results on three classifiers using 25 markers

Still based on previous research (Chapter 4), 14 'important' markers candidates were selected. The strategy was to select those markers that were discriminant in the categorisation process and whose distribution was very different among the six classes. These 14 markers were selected on the basis of clinical importance as indicated by pathologists involved in previous studies. An exhaustive search of the best combination of 10 markers out of these 14 was then performed based on the naive Bayes classification results. This was done as reducing the number of markers used for classification is a clinical aim, as this would both simplify and reduce the costs of a clinical test based on these markers.

This 'new' smaller dataset was used to repeat previous experiments applying the above classifiers on it. For the C4.5 decision tree a particular difference could not be seen, having 581 cases (87.6%) correctly classified. Also for the Multilayer Perceptron an increased number of misclassified instances was obtained, this time being 34 (5.1%). The naive Bayes, instead, performed very well compared to the previous run. Now 617 cases (93.1%) were classified properly and just 46 (6.9%) were misclassified.

A summary of the latter results is reported in Table 5.2.

Ten Markers

Method	*Classified*	*Misclassified*
C4.5	581 (87.6%)	82 (12.4%)
MLP	629 (94.9%)	34 (5.1%)
NB	617 (93.1%)	46 (6.9%)

Table 5.2: Comparison of results on three classifiers using only 10 markers

The 10 accuracies of each algorithm were compared using t-tests, after checking for normality using the Shapiro test [152]. It was found that, for both the whole data and the 10-markers datasets, the Multilayer Perceptron classifier performed significantly better than the other two ($p << 0.01$). The C4.5 decision tree algorithm was significantly more accurate than the naive Bayes ($p < 0.01$) when the whole data was considered, but was not when the number of features was reduced. Table 5.3 summarizes these findings.

	Average accuracies		
	C4.5	*MLP*	*NB*
Whole data	87.8 (6.3)	97.6 (1.8)	86.9 (2.5)
10 Markers	87.6 (6.6)	94.9 (2.6)	93.1 (2.5)

Table 5.3: Average accuracies on 10×10 cross validation experiments for the three classifiers (standard deviation in brackets)

5.4 Derivation of a new algorithm

From the results, all classifiers achieved a reasonable performance. They all are suitable for large-scale prediction and classification tasks on complex datasets. However, each of them has weaknesses. The C4.5 classifier may be considered what is called 'a white box model': the reason for arriving at the classification can be explicitly determined by examining the model. It also achieves good classification accuracy with large data in a short time. On the other hand, for real world datasets, the decision tree may become huge. In particular, for scientists not familiar with computational analysis, the set of rules coming from a decision tree may not be straightforward. Multilayer Perceptrons, using a back-propagation algorithm, are a standard algorithm for any supervised-learning pattern recognition process. However, like the majority of neural networks, it is a good example of a 'black box model', since explanation of the results is not available in an easily comprehended form. If one tries to write down the network model and the function representing the entire process, this might take a long time and in some cases it might be extremely complicated. Naive Bayes is a fast-supervised classification technique and, in general, it is a good approach for a classification problem. It is easy to understand and reproduce manually, being basically based on a product of conditional probabilities. However, one must be aware that naive Bayes relies on two fundamental assumptions: the first one is the complete independence of features (which is largely satisfied in the data under investigation), and the second is that the attributes should follow a normal distribution, which is not always true. Considering the latter assumption, it is immediately apparent that the 'Nottingham data' do not have a normal distribution. However, despite

the violation in its assumptions, the naive Bayesian classifier is remarkably effective on the dataset considered, showing a good performance.

Given the violation of the naive Bayes hypothesis of normality, other methods to represent features' distributions and to classify data were explored. A 'non-parametric' version of the naive Bayes classifier was implemented, with the aim of being able to categorize instances independently from their distribution.

5.4.1 A 'non-parametric' Bayesian classifier

The main idea of the new algorithm is that the closer a variable value is to its median in a particular class, the higher is the probability to be assigned to that specific group.

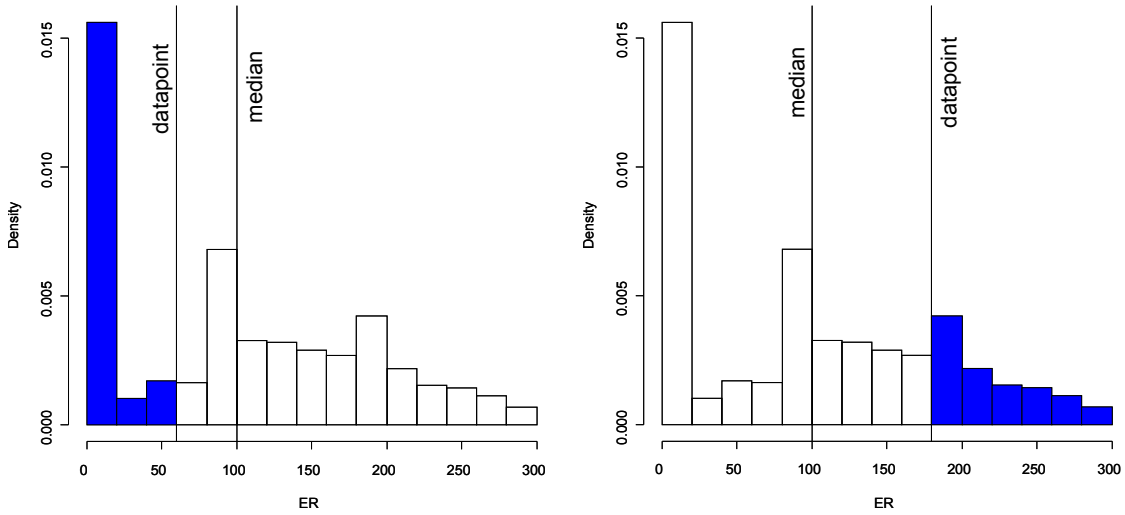
At the beginning of the algorithm the median value of each feature in every class was computed as well as the *priors* probabilities, which were defined as the ratio between each class numerosity and the total number of cases.

The following step is the main part of the method in which the single probabilities are calculated.

For each variable, it is checked whether the single variables' values are smaller or bigger than the median of that variable distribution in each class. If the value is smaller, the area under the histogram which remains on the left with respect to the value being analysed is calculated (Figure 5.4a). If the amount is bigger, the area on the right side is computed, taking in consideration the portion of the histogram delimited by the value and the maximum (Figure 5.4b). To compute the area under the histogram, the sum of each bar's area should be taken, and the latter has to be calculated as the product of bar's width and height. The amount returned is then divided by half of the total observations, as it is assumed that the total area under the histogram is equal to one.

In the next step, for each patient and each class, the product of all the features probabilities times the *priors* ones is computed.

$$p[i,k] = priors[k] \times \prod_{j=1}^{p} prob[j,k] \quad \text{for } k = 1,\ldots,K$$



(a) If the value is smaller than the median (b) If the value is bigger than the median

Figure 5.4: Area under the histogram

where j runs over the p variables, i represents patients and K is the number of groups.

The final step of the algorithm is the calculation of the prediction for each instance: it is defined as the class number which gives the highest $p[i,k]$ ($\arg\max_k p[i,k]$).

With a little abuse of notation, this proposed algorithm can be summarised in the following way: calling m, min and Max the median, minimum and maximum values of each feature in each class, and, for similarity with the naive Bayes, $g(x,m)$ the function that represents each variable distribution, the value of k for which $p[i,k]$ is maximised has to be found, where

$$p[i,k] = priors[k] \times \begin{cases} \prod_j \frac{1}{N/2} \int_{min}^{x} g(x;m) & x < m \\ \prod_j \frac{1}{N/2} \int_{x}^{Max} g(x;m) & x > m \end{cases},$$

j represents one of the features, x is the particular variable's value under investigation and i runs over the instances set.

In the following, the acronyms NB (naive Bayes) and NPBC (non-parametric Bayesian classifier) will be used to indicate, respectively, the usual naive Bayes classifier and

the new method.

5.4.2 Data sets considered

`Breast Cancer` Data Set

To validate the algorithm just described, the original 25 well-characterised biomarkers from Abd El-Rehim *et al.* data [1] will be used. In this data set, several clinical information were also available, including the Nottingham Prognostic Index (NPI) score and its defining factors (tumour size, grade, and stage of disease). The NPI was defined in [64] as a prognostic index which may be used to categorise patients affected by breast cancer according to its value. In particular, the index is calculated according to the following formula:

$$NPI\ Score = (0.2 \times size) + grade + stage$$

and five different groups may be defined depending on its value.

NPI Score	Prognostic Group
≤ 2.4	Excellent Prognostic Group (EPG)
2.5 - 3.4	Good Prognostic Group (GPG)
3.5 - 4.4	Moderate Prognostic Group 1 (MPG1)
4.5 - 5.4	Moderate Prognostic Group 2 (MPG2)
> 5.4	Poor Prognostic Group (PPG)

The 25 biomarkers were used to predict the NPI groups.

UCI Machine Learning Repository Data Sets

The other datasets used to validate the new method were taken from the UCI machine learning repository [8]: `vehicle`, `glass`, and `haberman`.

For the `vehicle` dataset, the purpose is to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. Each vehicle may be viewed from one of many different angles. The original purpose was to find a method of distinguishing 3D objects within a 2D image by application of an ensemble of shape

feature extractors to the 2D silhouettes of the objects. Measures of shape features extracted from example silhouettes of objects to be discriminated were used to generate a classification rule tree by means of computer induction. This object recognition strategy was successfully used to discriminate between silhouettes of model cars, vans and buses viewed from constrained elevation but all angles of rotation. The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System), which extracts a combination of scale independent features utilising both classical moments based measures such as scaled variance, skewness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness. Four 'Corgie' model vehicles were used for the experiment: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars [156].

The `glass` dataset is taken from the USA Forensic Science Service and six types of glass, defined in terms of their oxide content (i.e. Na, Fe, K, etc), are considered. The study of classification of types of glass was motivated by criminological investigation: at the scene of the crime, the glass left can be used as evidence, if it is correctly identified [8, 54].

The `haberman` dataset contains cases from study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer [8, 73].

Table 5.4 gives the description of the three UCI benchmark problems. Each variable in each dataset was tested for normality using the Shapiro test [152]. Since no test data sets were provided in the benchmark sets, the ten-fold cross validation option was again used to evaluate the performance of the proposed algorithm. That is, each dataset was split randomly into ten subsets and one of those sets was reserved as a test set; this process was repeated ten times and the mean of the results was used.

Since two of the UCI datasets (namely `vehicle` and `glass`) had been also analysed by

name	vehicles	glass	haberman
#pts	846	214	306
#ats	18	9	3
#cls	4	6	2

#pts: the number of training data;
#ats: the number of attributes of patterns;
#cls: the number of classes.

Table 5.4: Three benchmark datasets from UCI

Bouckaert [17], comparing three main methods for dealing with continuous variables in naive Bayes classifiers, a comparison with Bouckaert's results will be performed simply by looking at the average accuracies reported in the original work [17]. As a matter of fact, in [17], the kernel method and the discretisation one have been used in comparison with the original naive Bayes. The *kernel method* approximates $P(X|C)$ for a continuous variable X (see Equations 2.30 and 2.31 in Section 2.6.3) by a sum of so called kernels, which are functions centered around data points [17, 91]. The *discretization method* [41] instead, first discretises the continuous variables into discrete ones, leaving a simpler problem without any continuous variables.

5.4.3 Results

First of all, four cases for which the NPI value was missing were deleted from the `breast cancer` data. Then, the experiments were started running the naive Bayes classifier in WEKA using the 10-fold cross validation option and evaluating the accuracy of the obtained classification simply by looking at the percentage of the correctly classified instances. Also when using the new method in R, the 10-fold cross validation option was utilised.

It was found, for the `breast cancer` dataset, that only 249 (23.2%) patients were correctly assigned to their particular class, while the remaining 823 (76.8%) were misclassified. For the Statlog `vehicle` dataset, instead, naive Bayes properly classified 381 instances (45.0% of the total amount), leaving 465 cases (55.0%) incorrectly assigned to their group. When considering the `glass` dataset, the algorithm correctly classified

almost just half of the cases (48.6% which corresponds to 104 data points), leaving the other half (110 cases, equal to 51.4%) not properly classified. For the last dataset analysed (`haberman`), naive Bayes assigned 229 patients (74.8%) to the proper group, and just 77 (25.2%) were misclassified.

With the new algorithm, a substantial improvement in the amount of cases that were correctly classified was obtained when considering the `breast cancer` dataset and more cases were also correctly assigned to their group when the UCI datasets were analysed. For the `breast cancer` data, the number of patients which were assigned to their original class was 416 (38.8%), and 656 (61.2%) were wrongly classified. For the `vehicle` data, NPBC was able to properly classify 503 cases (59.5%), 122 more than with the naive Bayes. The remaining 343 instances (40.5%) were misclassified even with the new algorithm. When moving to the `glass` dataset, it was obtained that 121 (56.5%) types were correctly assigned to their group, while the remaining 93 (43.5%) were not. The last dataset considered, `haberman`, had 240 (78.4%) data points properly classified and 66 (21.6%) misclassified.

When using a multinomial Logistic Regression model (MLR) the number of cases correctly classified was higher with respect to previous techniques for the `vehicle` and `glass` data sets, but not for the `breast cancer` one. Concerning the `haberman` data, for which a generalised linear model (GLM) was fitted, the same results obtained with the naive Bayes classifier were returned: a total of 229 (74.8%) patients were correctly assigned to their class, while the remaining 77 (25.2%) were not.

A summary of the results is reported in Tables 5.5 to 5.8. In Tables 5.6 and 5.7 average accuracies of different naive Bayes methods computed by Bouckaert [17] are also reported.

Calibration plots (see Section 5.2.1) for the `breast cancer`, `vehicle` and `glass` data sets are reported in Figures 5.5, 5.6, and 5.7. If the curve is above the diagonal line (which represents a perfectly calibrated classifier), it means that there is an overestimation of the class membership probabilities by the classifier. It can be seen that,

`breast cancer` data

Method	*Classified*	*Misclassified*
NB	249 (23.2%)	823 (76.8%)
NPBC	416 (38.8%)	656 (61.2%)
MLR	332 (31.0%)	740 (69.0%)

Table 5.5: Comparison of results over the `breast cancer` dataset. NB: Naive Bayes, NPBC: Non-Parametric Bayesian Classifier, MLR: Multinomial Logistic Regression

`vehicle` data

Method	*Classified*	*Misclassified*
NB	381 (45.0%)	465 (55.0%)
NPBC	503 (59.5%)	343 (40.5%)
MLR	678 (80.1%)	168 (19.9%)
BK [17]	(60.9%)	(39.1%)
BD [17]	(61.1%)	(38.9%)

Table 5.6: Comparison of results over the Statlog `vehicle` dataset. BK: Bouckaert's Kernel method, BD: Bouckaert's Discretisation method

`glass` data

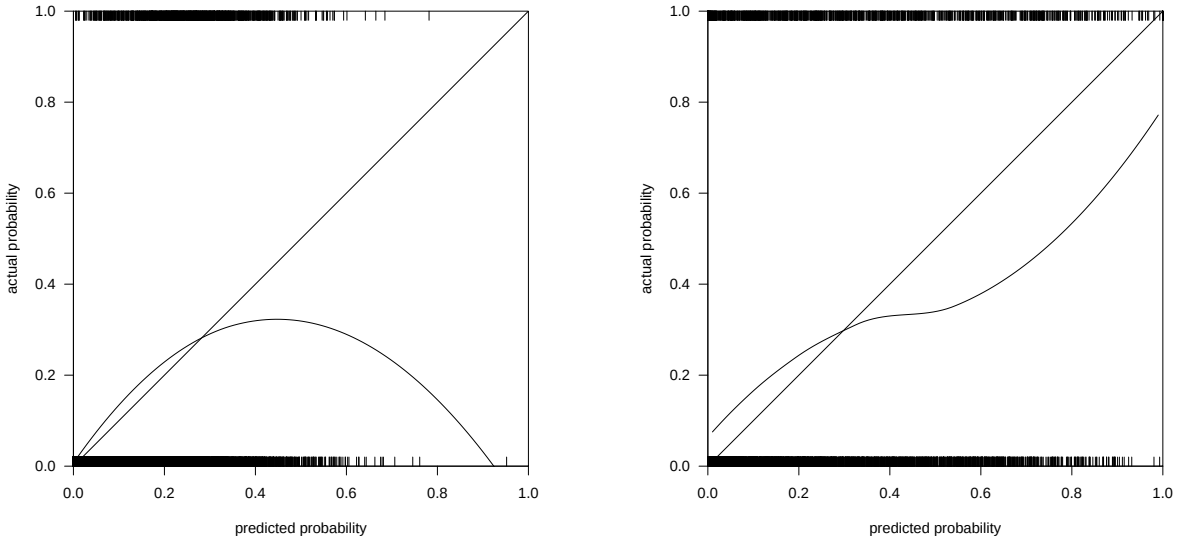
Method	*Classified*	*Misclassified*
NB	104 (48.6%)	110 (51.4%)
NPBC	121 (56.5%)	93 (43.5%)
MLR	134 (62.6%)	80 (37.4%)
BK [17]	(51.1%)	(48.9%)
BD [17]	(71.9%)	(28.1%)

Table 5.7: Comparison of results over the `glass` dataset

`haberman` data

Method	*Classified*	*Misclassified*
NB	229 (74.8%)	77 (25.2%)
NPBC	240 (78.4%)	66 (21.6%)
GLM	229 (74.8%)	77 (25.2%)

Table 5.8: Comparison of results over the `haberman` survival dataset. GLM: Generalised Linear Model

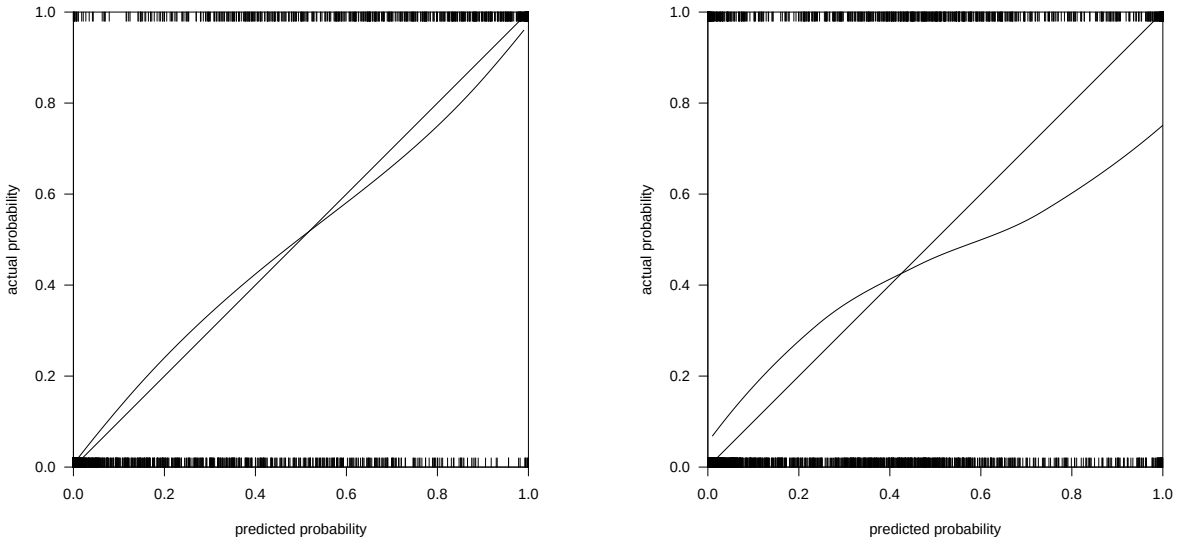


(a) MLR (b) NPBC

Figure 5.5: Calibration plots for multinomial logistic fit to the `breast cancer` data

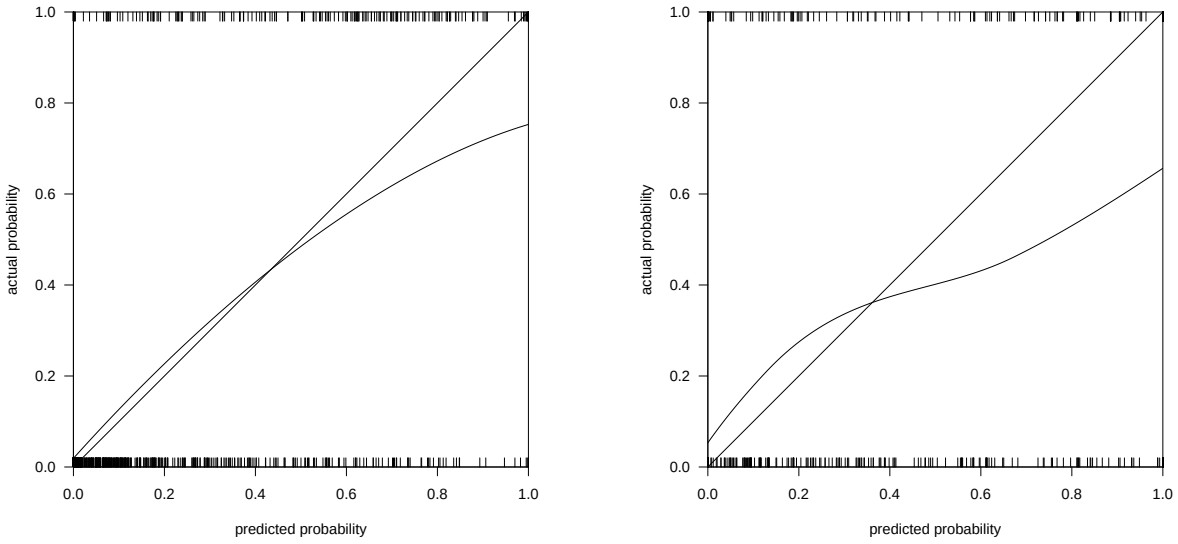
for the `breast cancer` dataset, the Logistic Regression model probabilities are less calibrated than the ones obtained by the proposed method. For the other data sets considered, Logistic Regression performed slightly better than the new algorithm.

For the `haberman` data set a plot of the ROC curves for both the GLM and the new method was produced and is reported in Figure 5.8. From the values of the areas under the curves, reported in the plot, a slightly better accuracy of the GLM is evident with respect to the new method, which, in any case, seems to be a quite good predictive model for the `haberman` data. Although similar plots could have also been produced when considering the other data sets, it was decided to use only those data where the response variable was



(a) MLR (b) NPBC

Figure 5.6: Calibration plots for multinomial logistic fit to the `vehicle` data



(a) MLR (b) NPBC

Figure 5.7: Calibration plots for multinomial logistic fit to the `glass` data

binary (0 or 1) and not multinomial.

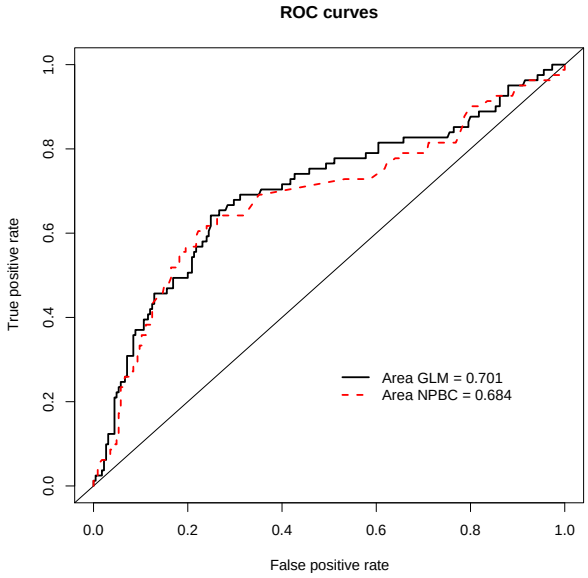


Figure 5.8: ROC curves for `haberman` survival data

5.5 Discussion of results

In the experiments presented in this chapter, several different results were obtained from the classifiers. Using the whole Nottingham 'in-class' dataset (25 markers $\times$ 663 instances), the best performance was obtained from the Multilayer Perceptron classifier: in fact just 16 cases were incorrectly classified. The naive Bayes and C4.5 decision tree returned similar results (but worse than the MLP), with the latter being 'only marginally' more accurate than the naive Bayes.

When just the 10 'most important' markers were considered, a substantial improvement in the naive Bayes performance was found: even though it did not return the highest number of correctly classified instances, it performed much better than with all the markers, decreasing the number of misclassified instances from 87 to 46. Again, the best results were obtained using the Multilayer Perceptron, but this time the network did not perform as well as before: there were 18 more cases of misclassification. Finally, the C4.5 decision tree was the worst classifier among the three used, performing almost identically as with all markers.

Starting from these results, a 'non-parametric' approach of the naive Bayes classifier

to deal with continuous and non-normal covariates was developed. The method was presented and its performance over four particular data sets was compared to the original naive Bayes one as well as to Logistic Regression.

The naive Bayes method did not perform well on all data considered in this part of the work, and, focusing on the `breast cancer` one, this reflects a sort of independence between biological markers and clinical information. Moreover, all datasets' features strongly violated the normality assumption, so proving the reason of a 'non-parametric' approach.

For each class and each variable, the median value and the histogram of its distribution were computed. Different situations that might occur were then considered: if, fixing a particular class and a particular data point, the value of a generic variable was lower or greater than the extreme values of the same variable in the class considered at that stage, then a probability close to zero to belong to the specified class was assigned to that data point; if the value was identical to the median the probability was set to be one; finally, if the data point was smaller than the median, the area between the distribution's minimum and the actual value was calculated (or between the value and the distribution's maximum if value was greater than the median). The value obtained was then divided by half number of observations. As for the naive Bayes classifier, it was calculated, for each case, the product of probabilities of all features given the classes. Data were classified looking at the class number for which the above reached the maximum.

With the method just described, a bigger amount of data points was correctly classified, raising the percentage from 23.2% to 38.8% for the `breast cancer` dataset, from 45% to almost 60% for the `vehicle` Statlog dataset, from 48.6% to 56.5% for the `glass` data, and from almost 75% to more than 78% for the `haberman` dataset. However, when using Logistic Regression, different results were obtained. For the `breast cancer` dataset, the proposed new model seemed to be more accurate (in terms of percentages of patients correctly classified) and more calibrated with respect to the Logistic Regression. This was not true when considering the UCI data sets, for which the new algorithm

slightly appeared to be less calibrated and less accurate. However, for the `haberman` dataset, when a GLM was fitted to the data, the number of patients correctly assigned to their class was identical to the one obtained when using naive Bayes and the ROC curve associated to the method presented here was very similar to the one produced by the GLM, providing two close values for the areas under the curve.

It is important to note that a couple of data sets presented in this work were also used in [17] to compare naive Bayes normal method with the kernel and the discretization ones obtaining both better and worse results compared to ours (Tables 5.6 and 5.7). Bouckaert considered those three methods to deal with continuous variables when using the naive Bayes classifier. Instead the 'non-parametric' method was developed to deal with the non-normality of several dataset variables and, moreover, it outperformed all the ones proposed in [17] when applied over the `breast cancer` dataset (results not reported).

It is also worth noting that the new developed method is not meant to be applicable over all available datasets. In this chapter several situations were presented, for which a classical approach, the naive Bayes classifier, was outperformed by a more general algorithm that does not assume any particular distribution of the analysed features. In general, according to the experience, the classical naive Bayes classifier outperforms the new method when datasets with categorical features are considered or when the majority of them follows a normal distribution. In these situations it is advisable to use the original naive Bayes approach.

5.6 Summary

In this chapter, supervised learning was applied over several case studies. In particular, three different classifiers, the C4.5 decision tree, the MLP-ANN and the naive Bayes were reviewed and used over the 'in-class' patients of the Abd El-Rehim *et al.* [1] breast cancer dataset in order to validate the previous classification derived and characterised in Chapter 4. Surprisingly, the naive Bayes classifiers performed quite well, especially when

just the 10 'most-important' markers were considered. This happened even though one of the underlying assumptions of the NB was strongly violated by the data: as a matter of fact, all the features did not follow a normal distribution. A non-parametric version of the naive Bayes was then developed and validated over known data sets. These latter results were presented in this chapter together with their comparison with the Logistic Regression approach.

In the next chapter, a novel clustering technique, called Affinity Propagation, will be presented. This algorithm combines properties of both heuristic and model-based approaches and it was shown to be feasible with very large data sets. Its application over several data set of cancer will be reported and discussed.

Chapter 6

The Affinity Propagation Method: Is It Computationally Efficient?

In this chapter, the clustering technique Affinity Propagation (AP, [60]) will be presented and adopted for grouping tumours with similar biological characteristics. This method combines properties of both hierarchical and model-based clustering. A particular feature of AP is that the number of clusters should not be passed as an input parameter. Instead, a measure of *preference*, as described later, has to be chosen. By iterating the algorithm over a set of *preferences*, it is possible to obtain an indication about the number of clusters present in the data.

The Affinity Propagation was analysed in this study to find out whether it could enhance the proposed framework (described in the next chapter) by using a different approach to discover the best classification in a dataset. In the next sections, the AP algorithm will be described in details, together with motivations for its use. Then, results obtained by applying this method to cancer case series will be reported. In the end, the evaluation of the computational complexity of AP and its comparison with the ones of 'traditional' algorithms will be presented to address the real efficiency of the algorithm over the case studies analysed.

6.1 Background and motivation

Genomic analysis renewed interest in clustering techniques. After the seminal paper of Eisen and colleagues [47], proposing hierarchical clustering and the visual inspection of the dendrogram to discover unknown pattern of gene associations, the use of clustering has become more and more popular especially to discover profiles in cancer with respect to high-throughput genomic data. Important applications of the Eisen method are the work of Bittner *et al.* [15] on clustering of cutaneous melanoma and the works of Perou *et al.* [137] and of van't Veer *et al.* [170] on breast cancer. More recently, a classification of breast carcinoma using traditional tumour markers was proposed [5]. Different clustering algorithms were used in [5] to choose a stable solution across different clustering methods. At last a classification in four clusters was preferred and suggested a possible separation of high risk profiles. This classification [5] was in agreement with those obtained with c-DNA microarray data [137, 170].

One of the main problems connected with cluster analysis is the choice of the number of clusters. The visual inspection of the dendrogram suggested by Eisen *et al.* [47] is an informal method to determine the number of clusters. Such a procedure was indeed criticized by Goldstein *et al.* [70] as it can cause difficulty in assessing the validity of the grouping. In classical cluster analysis it is customary to use indexes to compare one cluster solutions to other cluster solutions and to choose the one suggested as optimal. In a recent study [5], different indexes were used to select an optimal partition. Namely the indexes proposed by Calinski and Harabasz [20], Krzanowski and Lai [102], Hartigan [77] and Tibshirani *et al.* [165], were considered. According to Getz *et al.* [68], the number of clusters should be determined internally by the clustering algorithm and should not be externally prescribed.

In this chapter, a recently developed clustering algorithm, the Affinity Propagation [60], will be used to cluster cancer patients in order to evaluate its performance with respect to the traditional approaches. Although this algorithm does not determine automatically the number of clusters, it provides a consistent method to suggest the number

of groups to be created which can be useful to detect different levels of association pattern. Several datasets, already published in literature were considered in this study: the melanoma data of Bittner *et al.* [15], and four different breast cancer data: namely, the studies of Ambrogi *et al.* [5], Perou *et al.* [137], van't Veer *et al.* [170] and the 'Nottingham dataset' presented in [1]. In addition, the computational complexity of AP was analysed and compared to the one of K-means. These two methods were applied over the 'Nottingham dataset', and a comparison of the required CPU times was performed in order to evaluate whether it was worth including AP in the proposed framework.

6.2 The Affinity Propagation algorithm

As other clustering algorithms, Affinity Propagation uses data to find a set of centres such that the sum of squared errors between data points and their nearest centre is small. Like other traditional clustering techniques, the Affinity Propagation algorithm determines the centres from real data points (which, for this method are called 'exemplars'). These exemplars correspond, for instance, to the medoids in the algorithm PAM [97] (Partitioning Around Medoids, a more robust version of K-means), that is k representative objects among the observations of the dataset that should represent the structure of the data. As a technical detail, it is worth noting that K-means algorithm does not use exemplars, as the centres are not generally actual data points but are computed as the means of data points belonging to clusters.

Affinity Propagation combines the properties of different classes of clustering algorithms. On one hand, algorithms like hierarchical clustering are based on grouping pairs of objects with high affinity. On the other hand model-based clustering uses a probability model based on a mixture of class conditional distributions. Affinity Propagation uses both pairs comparison and a probability model to determine the optimal grouping. According to a more technical point of view, Affinity Propagation can be derived as the sum-product algorithm in a graphical model describing the mixture model [59].

AP is a method that recursively transmits messages (that will be defined subsequently) between pairs of data points until a good set of exemplars and corresponding clusters emerges.

The first step for the algorithm implementation is to choose a measure of *similarity*, $s(i,k)$, between all pairs of data points. In AP terminology, $s(i,k)$ quantifies how well the data point with index k is suited to be the exemplar for data point i. In cluster analysis, the negative Euclidean distance is generally used as a similarity measure. When dealing with c-DNA data, however, it is common to use the Pearson correlation [15]. The same choices can be made when using Affinity Propagation [60].

Rather than requiring the number of clusters to be prespecified, AP assigns a common value (called *preference*) to all data points. The second step relates to the choice of these values of preference which are indicated as $p(i)$. The preferences represent a measure of how much data point i is candidate to be an exemplar. In general, data points with larger values of $p(i)$ are more likely to be chosen as exemplars. At the beginning, the AP simultaneously considers all data points as potential exemplars (so the preferences being the same for all data points). The number of identified exemplars is influenced by the values of the input preferences, but also emerges as a result of the message passing structure that is illustrated subsequently. For very small value of input $p(i)$, for every i, all data points are grouped in one large cluster with a single exemplar; in the opposite case of large $p(i)$ for every i, each data point prefers to be its own exemplar. In general, the initial value of the preferences is set equal to the median of all input similarities (resulting in a moderate number of clusters) or to their minimum (resulting in a small number of clusters).

The algorithm is named Affinity Propagation because at any point in time, each message reflects the current affinity between a data point and its exemplars. In practice, during the message-passing algorithm, each data point i furnishes a measure to suggest another data point k to be selected as cluster centre, taking into account other potential exemplars for point i. There are two kinds of message being passed between each pair of data points

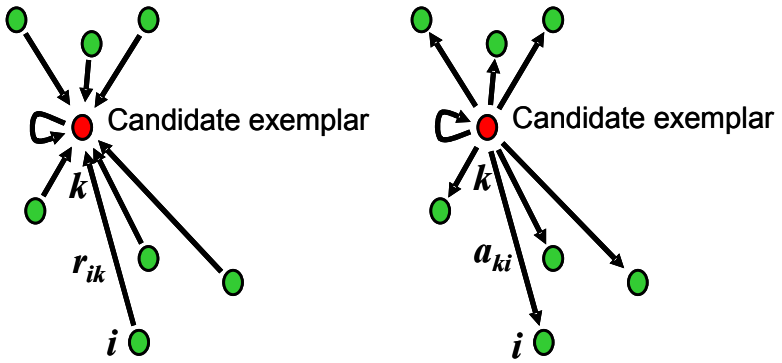


Figure 6.1: Message exchange between data points

that represent the relationship between data points:

- "responsibility": sent from data point i to candidate exemplar k. It is a measure that quantifies how well-suited point k is to be the exemplar for point i, taking into account other potential exemplars for point i. This message is represented by $r(i,k)$ and is computed using this formula:

$$r(i,k) = s(i,k) - \max_{k':\ k' \neq k} \{a(i,k') + s(i,k')\};$$

- "availability": sent from candidate exemplar point k to point i. It is a measure that reflects the evidence for point i to choose point k as its exemplar, considered that other points may have k as an exemplar. This message is represented by $a(i,k)$ and is computed using this formula:

$$a(i,k) = \min\{0, r(k,k) + \sum_{i':\ i' \notin \{i,k\}} \max\{0, r(i',k)\}\}$$

An illustration of the responsibilities and availabilities is reported in Figure 6.1. A particular measure is the "self-responsibility", that is $r(k,k)$; it reflects accumulated evidence that point k is an exemplar and how it would be unsuitable to be integrated in a group of another cluster centre.

At the beginning of the algorithm, the availabilities are initialized to zero, so $r(i,k)$ is

set to the input similarity between point i and its potential exemplar k minus the largest of the similarities between point i and other candidate exemplars. After the computation of all the responsibilities, the availabilities are worked out using the previous formula. Only the positive portions of responsibilities between the candidate exemplar k and other data points i' are added because it is only necessary for a good exemplar to explain some data points well ($r(i',k) > 0$) regardless of how poorly it explains other data points ($r(i',k) < 0$). In fact, if $r(i',k) < 0$, k is not suited to be the exemplar for point i'. So in this case, the point i' will not contribute to the message passing from candidate exemplar k to point i.

After that, the messages are recursively updated for a fixed number of iterations or until a stable clustering result. At any stage, the availabilities and responsibilities can be combined to identify exemplars. For point i, the value k that maximises $a(i,k) + r(i,k)$ identifies point i as exemplar if $k = i$ or identifies the data point that is the exemplars for point i. In other words, as suggested in [43], after exchanging messages, Affinity Propagation identifies a set of exemplars K so as to maximise the net similarity, which is defined as

$$\sum_{i \notin K} \max_{k \in K} s(i,k) + \sum_{k \in K} p(k)$$

where $p(k)$ is the *a priori* preference that point k would be chosen as an exemplar. At the end of the message passing, the number of clusters is obtained together with the labels for each data point of its exemplars. All the equations shown above are derived and explained in detail in the Supporting Online Material for [60].

An illustration of how Affinity Propagation works is reported in Figure 6.2. Affinity Propagation is illustrated for two-dimensional data points, where negative Euclidean distance was used to measure similarity. Each point is coloured according to the current evidence that it is a cluster centre (exemplar). The darkness of the arrow corresponds to the strength of the transmitted message.

AP differs from other clustering algorithm like K-means in several aspects. Firstly, K-means is based on the minimisation of the Euclidean distance between data points and

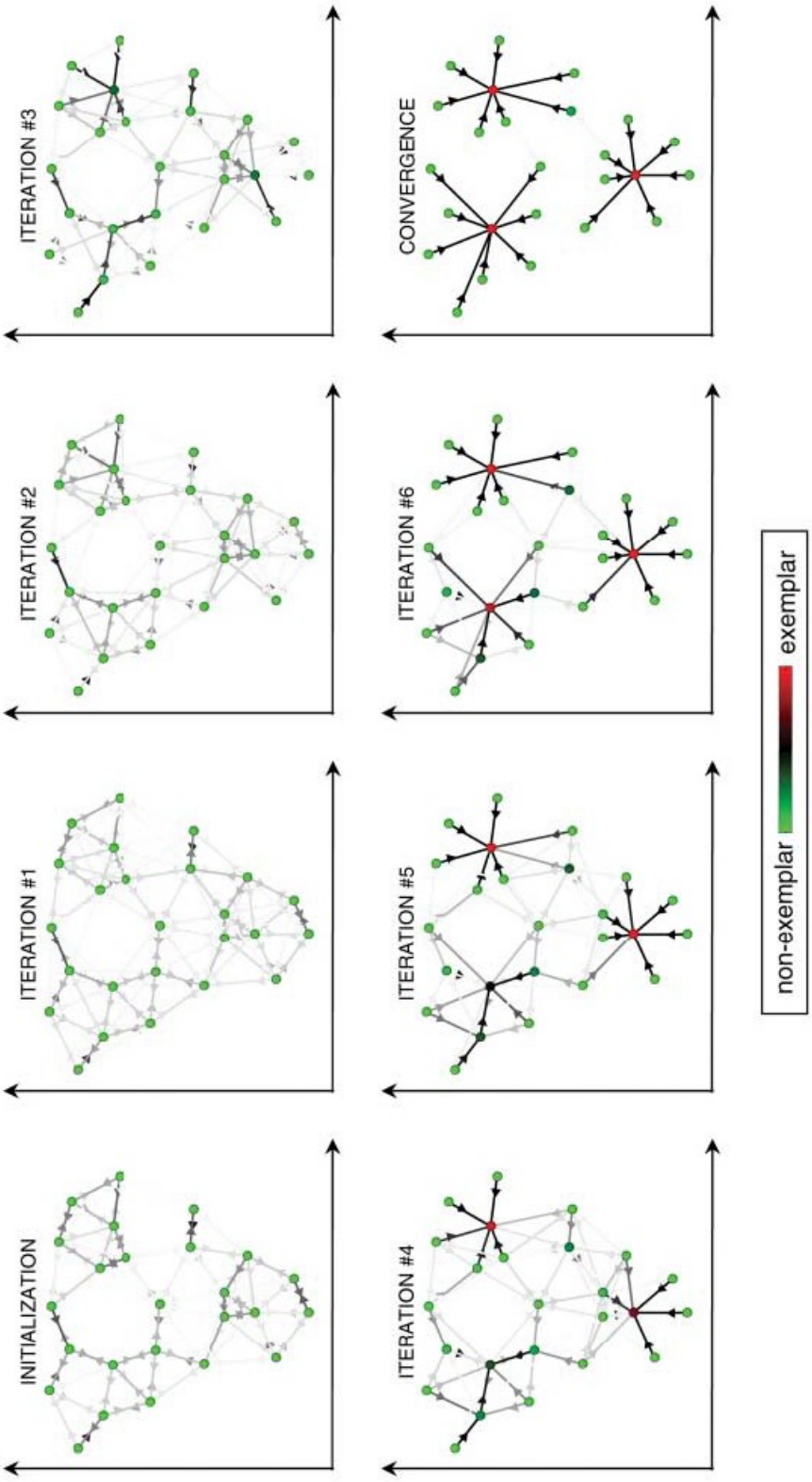


Figure 6.2: How Affinity Propagation works [60]

cluster centres, which are computed as mean points within a cluster, while AP uses general dissimilarities and actual data points as centres. Secondly, K-means begins with an initial set of randomly selected centroids and iteratively refines this set so as to decrease the sum of squared error. AP, instead, simultaneously considers all data points as potential exemplars [60]. Finally, the use of squared Euclidean distance as measure of dissimilarity between data points and centroids in K-means not only limits the type of data variables that can be considered, but it can also make the determination of the cluster means non-robust to outliers [14]. Furthermore, as the K-means algorithm can be thought as an EM algorithm, it is worth noting that the use of a general dissimilarity measure in such a context implies an increased complexity in the maximisation step which is generally faced by considering as possible cluster centres actual data points. By doing so, the algorithm can be implemented for any choice of similarity measure [14].

An advanced characteristic of Affinity Propagation is that it determines the number of clusters on the basis of the message passing architecture and of the points that are most representative, given an initial common preference. It is possible to see the effect of the value of the input preference on the number of clusters by a graphic with the value of the common initial preference on the x-axis and the respective number of clusters on the y-axis. In this way, the value to adopt in the analysis can be established in correspondence with plateaus that are observable in this graphic. Given an initial preference, AP defines a unique solution.

One of the strong points of AP is its computational efficiency, as described in [109]. Leone and colleagues, indeed, state that AP seeks at maximising the overall similarity of all data points to their exemplars under the hard constraint which forces each data point in a cluster to refer to its exemplar and each exemplar to refer to itself as a self-exemplar. The solution of this hard combinatorial task is approximated following the ideas of belief-propagation [103]. Whereas an implementation of belief propagation for n data points leads to $O(n^3)$ messages which have to be determined self-consistently, the formulation of Frey and Dueck [60] allows to work with $O(n^2)$ messages only. There-

fore, the algorithm is feasible even for very large data sets [109], such as a collection of 75,066 segments of DNA (60 bases long) corresponding to putative exons [60]. This last assumption was actually also confirmed in the original work of Frey and Dueck [60], where the authors compared the AP performance on four different scenarios with the one of the K-means algorithm. When dealing with the problem of clustering images of faces using standard optimisation criterion of squared-error, Frey and Dueck used both AP and K-means to identify exemplars among 900 grayscale images extracted from the Olivetti face database[1]. The authors claimed that AP found exemplars with much lower squared error than the best of 100 runs of K-means clustering, which took about the same amount of computer time [60].

Considering the time complexity of the affinity propagation, Zhang *et al.* [196] stated that, while the message passing algorithm converges with $n \log n$ complexity, the similarity matrix is computed with quadratic complexity, thus hindering the scalability of the approach. As a matter of fact, in Frey and Dueck [60] the similarity matrix was assumed to be given beforehand, or to involve a small fraction of the item pairs [196]. Algorithms like K-means or PAM, instead have an overall complexity of $O(nkt)$ or $O(tk(n-k)^2)$ (where n is the dimension of the dataset, k is the number of clusters and t the number of iterations), as already reported in Table 2.1 of Chapter 2.

6.3 Application of AP over known datasets

The Affinity Propagation was applied over several case studies of both breast and cutaneous cancers. In the following, each dataset will be presented in detail and subsequently results obtained by AP will be compared with those presented in the original papers. Finally, a comparison between the CPU times required to perform AP and K-means algorithms is reported.

[1]The Olivetti dataset, along with the similarities used to obtain the results described in [60] are available at www.psi.toronto.edu/affinitypropagation.

6.3.1 Case studies considered

This section includes a description of the case series considered in the work.

1. The information on 633 patients operated on for primary infiltrating breast cancer between 1983 and 1992, archived at the Pathology department of the University of Ferrara, was retrospectively analysed in the work of Ambrogi *et al.* [5]. The available data concerned patient age, pathological tumour size, histologic type, pathologic stage, and number of metastatic axillary lymph nodes; as well as immunohistological determinations of oestrogen receptor status (ER), progesterone receptors status (PgR), Ki-67/MIB-1 proliferation index (MIB1), *c-Erb*B-2/NEU (HER2) and the p53 oncosuppressor gene (p53). Values of ER, PgR and HER2 tended to be grouped on the following values: 0%, 10%, 25%, 50%, 75% and 100%; they were consequently discretized on those values. Values of MIB1 and p53 were used as originally measured.

2. The melanoma dataset of Bittner *et al.* [15] was also analysed. These data consist of gene expression profiles obtained on a collection of 38 samples, comprised of 31 melanoma tumours and 7 controls. For the analysis described in Section 6.3.2, the data from the seven control specimens were excluded and only the ratios for the 3613 genes that were considered 'well measured' (that is their intensities were sufficiently high) were used. These ratios were converted to log2 ratios.

3. Another dataset on breast cancer that was considered for investigation was that of Perou *et al.* [137]. Variation in gene expression patterns in a set of 65 surgical specimens of human breast tumours from 42 different individuals have been characterised, using complementary DNA representing 8102 human genes. According to the authors, these patterns provided a distinctive molecular portrait of each tumour [137]. Sets of co-expressed genes were identified for which variation in messenger RNA levels could be related to specific features of physiological variation. The tumours could be classified into subtypes distinguished by differences in their

gene expression patterns. In their paper, Perou *et al.* focused on a set of 1,753 genes in 65 experimental tissue samples. In each sample, the ratio of the abundance of transcripts of each gene to its median abundance across all tissue samples, is reported. Data and original analyses of both Bittner and Perou studies are fully described in the book "Design and Analysis of DNA Microarray Investigations" by Simon and colleagues [157].

4. The dataset analysed in van't Veer *et al.* [170] was also considered in this study: they used DNA microarray analysis on primary breast tumours of 117 patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases in patients without tumour cells in local lymph nodes at diagnosis. All patients were lymph node negative and under 55 years of age at diagnosis. Unsupervised clustering detected two subgroups of breast cancer, which differ in ER status and lymphocytic infiltration.

5. The last case study considered was the data set originally analysed in Abd El-Rehim *et al.* [1]. As already mentioned, these data consist of tumours of an independent set of 1076 patients from the Nottingham Tenovus Primary Breast Carcinoma Series, which were evaluated by Immunohistochemistry (IHC) for 25 markers. Levels of IHC reactivity were categorized using a modified H-score (values between 0-300). Further description of these data can be found in [1] as well as in previous chapters of this thesis.

6.3.2 Results

In this section, results obtained by the application of AP over known case studies will be reported in order to compare them with those obtained by hierarchical or partitional algorithms. Al last, the CPU time needed for the computation of affinity propagation over a single dataset was evaluated. Its comparison with the time requested by 'traditional' methods will be presented and discussed in order to evaluate the inclusion of this novel

algorithm in the proposed framework for the elucidation of core classes in a dataset.

Ambrogi *et al.* breast cancer biomarkers data

AP was applied to the breast cancer data analysed in [5] in which the final clustering was obtained using a K-Medoids algorithm to generate four clusters. The negative Euclidean distance was chosen as the similarity measure, in accordance with the original paper [5]. A graphical evaluation of the effect of the value of *preference* on the number of clusters for the breast cancer data is reported in Figure 6.3. The presence of three main plateaus in correspondence with two, four and five clusters is shown.

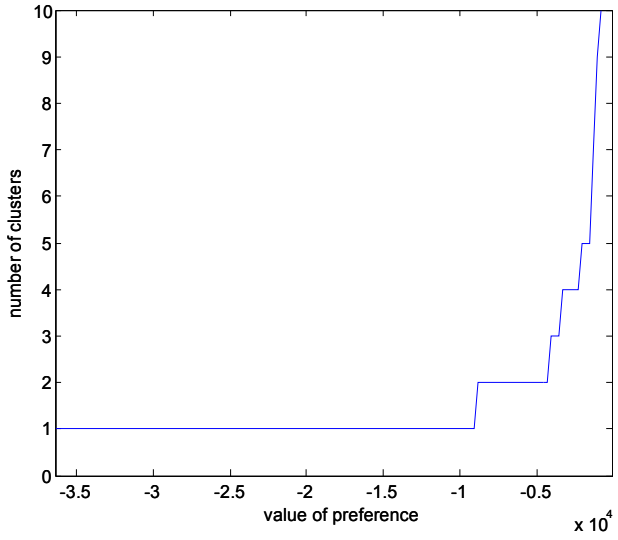
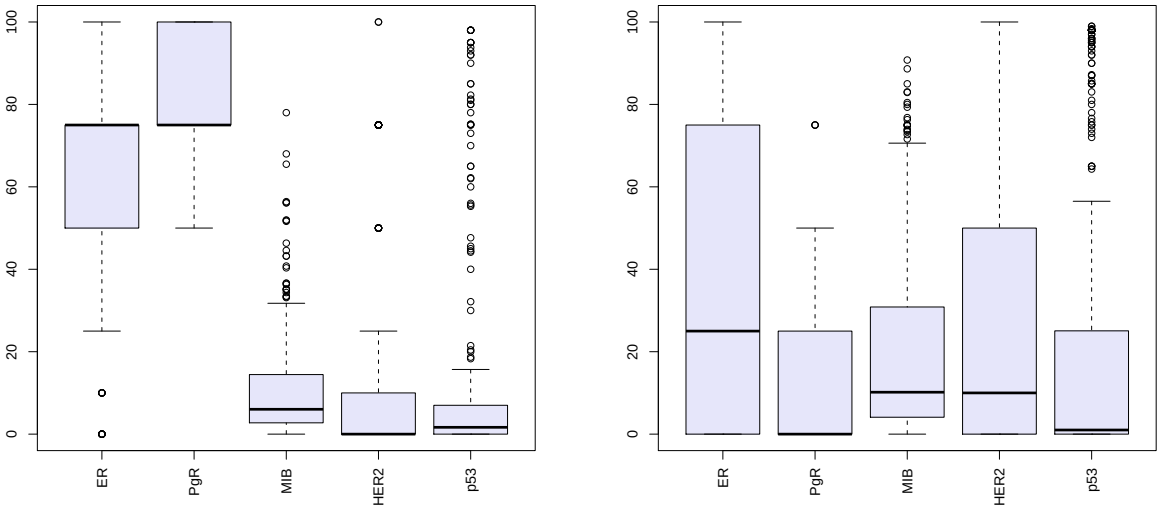


Figure 6.3: The effect of the value of the input preference on the number of clusters for the Ambrogi *et al.* data

When doing the analysis with two clusters, by using an input preference value in correspondence with that plateau, results consistent with knowledge from the literature were obtained. Indeed, one cluster was associated with low values of ER and PgR, while the other with high values of these biological markers, as reported in Figure 6.4.

The message-passing algorithm was then run with an input preference to obtain 4 clusters. The results are reported in Figure 6.5. From these plots, the resulting clusters could be characterised as follows. Cluster 1 was associated with highest values of ER and



(a) Distribution in AP cluster 1

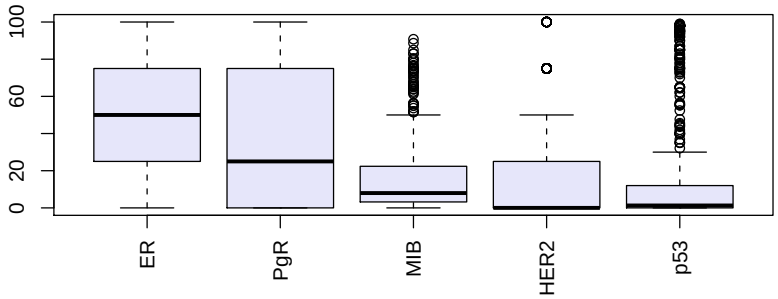
(b) Distribution in AP cluster 2

Figure 6.4: Distribution of variables in two clusters for the Ambrogi *et al.* data

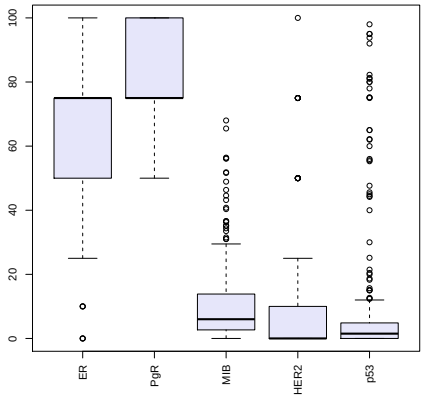
PgR, so it seemed to represent subject with characteristics known to be good prognostic factors. Cluster 2 seemed to be associated with intermediated values of PgR and ER and null values of HER2; so also this cluster was associated with less aggressive tumour features. Null values of ER, highest values of p53, HER2 and MIB1 were associated with Cluster 3. Null values of PgR and high values of HER2 were associated with Cluster 4. Therefore, Cluster 3 and Cluster 4 represent groups that are associated with characteristics known to be poor prognostic factors. As for the triple negative patients, null values of PgR, ER and HER2 associated with positive values of p53 were grouped in Cluster 3.

The distribution of subject between the classification using K-Medoids (PAM) and the classification using AP is reported in Table 6.1. If these results were compared with those of the previous work, null values of PgR, ER, HER2 and p53 were grouped in original Cluster 2, which was the cluster most similar to the characteristics of total sample. Instead, in this new classification null values of PgR, ER and HER2 associated with null values of p53 lay in Cluster 4, a cluster that is not similar to total sample for the distribution of biological markers and represents groups with poor prognostic factors.

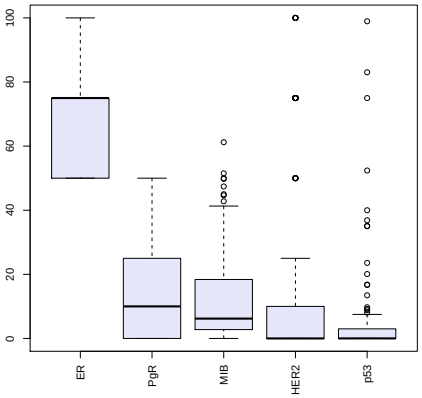
Afterwards, the AP algorithm was applied again to obtain a division of subjects in 5



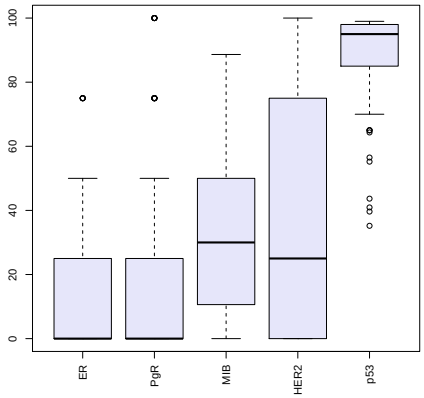
(a) Distribution of the whole data



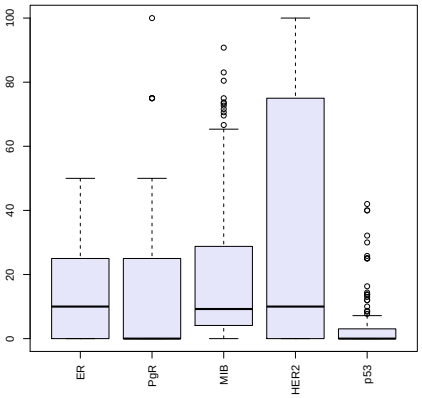
(b) Distribution in AP cluster 1



(c) Distribution in AP cluster 2



(d) Distribution in AP cluster 3



(e) Distribution in AP cluster 4

Figure 6.5: Boxplots for the whole data and grouped by AP cluster for the Ambrogi *et al.* data

		AP CLUSTERS			
PREVIOUS WORK'S CLUSTERS		1	2	3	4
	High ER	253	1	0	2
	Intermediate ER	1	122	0	84
	Low ER / High p53	0	1	88	2
	Low ER / High HER2	1	25	1	52

Table 6.1: The distribution of subjects between new and old classification (Ambrogi *et al.* data)

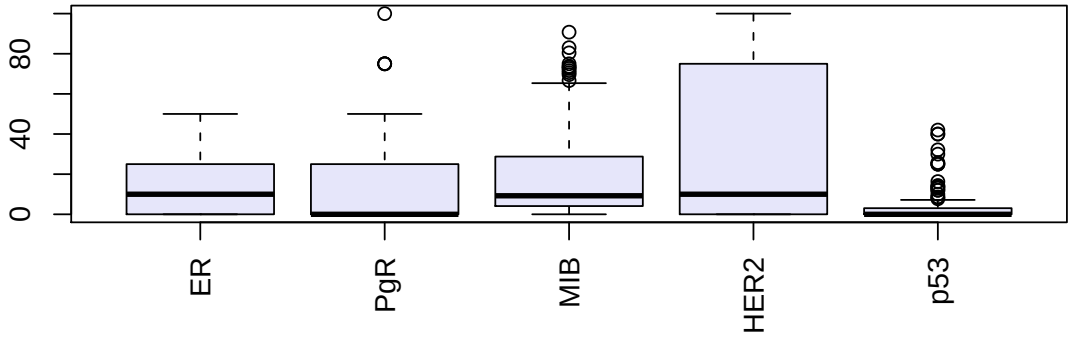
groups and to compare these results with the four clusters obtained in the previous run. This comparison is reported in Table 6.2, from where it is visible that the solution with five clusters is very similar with the one with four groups. There is just a subdivision in groups 4 and 5 of the last group determined in the previous run. However, this distinction into two separate groups is somehow significant, as it is shown in Figure 6.6. (a) shows the distribution of the five markers in cluster 4 for the AP with four groups, while (b) and (c), respectively, show the distributions of markers in clusters 4 and 5 for the AP solution with five groups. When the dataset is divided in five clusters, the fourth and the fifth differ by the presence or absence of HER2.

		AP 5 CLUSTERS				
AP 4 CLUSTERS		1	2	3	4	5
	1	210	41	0	2	2
	2	1	125	0	23	0
	3	0	0	86	1	2
	4	1	0	1	49	89

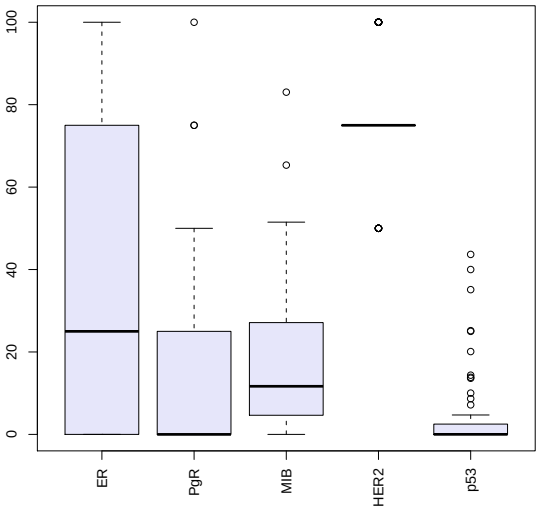
Table 6.2: The distribution of subjects between AP 4 and 5 groups (Ambrogi *et al.* data)

Bittner *et al.* melanoma data

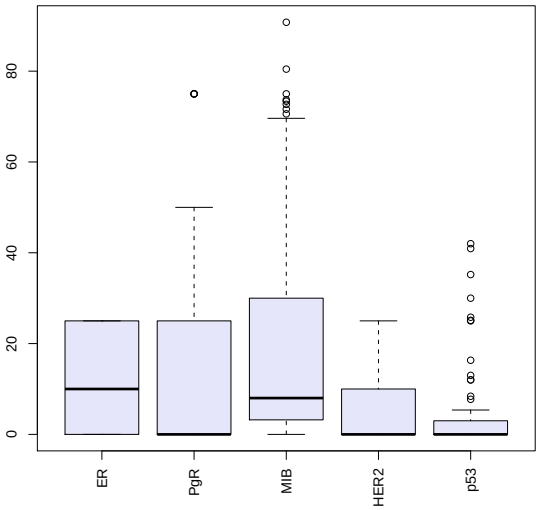
Bittner and colleagues [15] attempted to determine if c-DNA microarray data could be used to identify distinct subtypes of cutaneous melanoma, a malignant neoplasm of the skin. In particular they were able to identify two major cancer profiles with different biological characteristics. Results were based on the application of a hierarchical algorithm and on cutting the dendrogram by visual inspection [70]. In Figure 6.7 the dendrogram resulting from the application of a hierarchical algorithm with average linkage and a sim-



(a) Cluster 4 in AP four groups



(b) Cluster 4 in AP five groups



(c) Cluster 5 in AP five groups

Figure 6.6: Boxplots of markers for different AP groups

ilarity matrix based on Pearson correlation is reported. The two clusters were obtained by cutting the tree as shown in Figure 6.7. In this way the 31 melanomas were divided in a single group comprising 20 melanomas while the remaining 11 (actually grouped in 4 clusters) were considered together.

AP algorithm was applied to the melanoma data using a distance matrix based on correlations. The resulting plot of the cluster number for different *preference* levels is reported in Figure 6.8. The plot suggests solutions with 2, 4 and 5 clusters. An interesting

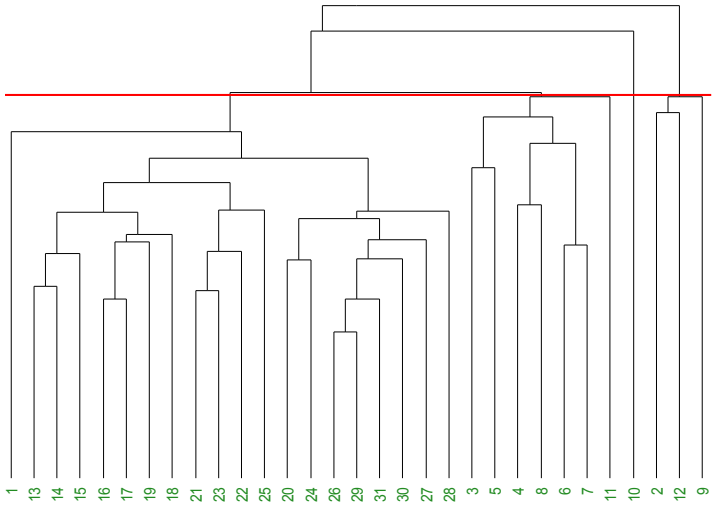


Figure 6.7: Dendrogram resulting from the application of hierarchical algorithm to Bittner *et al.* dataset

issue which can be raised from this plot is related to the correspondence between the number of clusters and the *preference* values. Frey and Dueck, in a FAQ webpage for Affinity Propagation[2], state that "the number of clusters is close to being monotonically related to the preference" but they do not investigate further on this. In Figure 6.8 it can be seen that, after the plateau at four clusters, the blue line decreases and then increases again to reach the following plateau at five clusters. Increasing the number of the possible values of *preference*, the plot reported in Figure 6.9 can be obtained. It can be seen that a small plateau is also found in correspondence of three clusters, while, if a fewer number of values of *preference* is considered, the relationship between the number of clusters and the *preferences* becomes monotonous (see Figure 6.10). This behaviour was not evident in any other similar plot reported in this chapter, thus leaving the issue of the monotonicity of this kind of plot open for future investigation.

The solution with 5 clusters is the one more similar to the one obtained by Bittner and colleagues. The 3-dimensional principal component plot in Figure 6.11 shows the two groups of the 31 melanomas. The red crosses correspond to the "interesting" cluster identified by Bittner and colleagues. The four black squares are tumours classified differently

[2]Available at http://www.psi.toronto.edu/affinitypropagation/faq.html

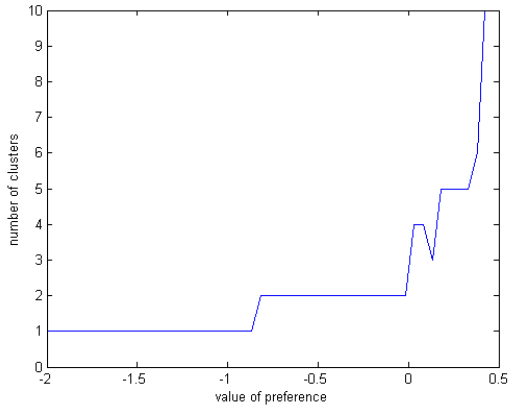


Figure 6.8: The effect of the value of input preference on the number of clusters for the melanoma data

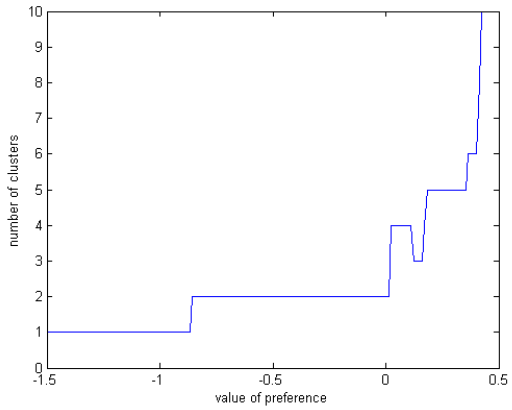


Figure 6.9: The effect of the value of input preference on the number of clusters for the melanoma data (‘zoom in’)

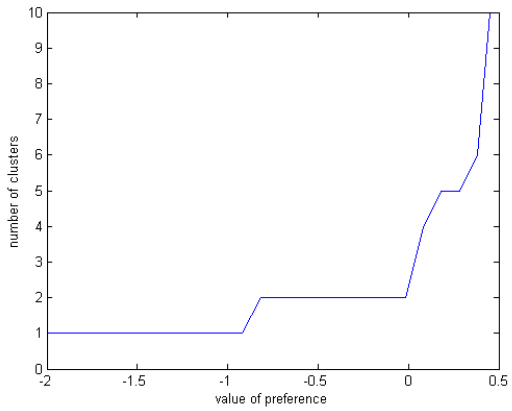


Figure 6.10: The effect of the value of input preference on the number of clusters for the melanoma data (‘zoom out’)

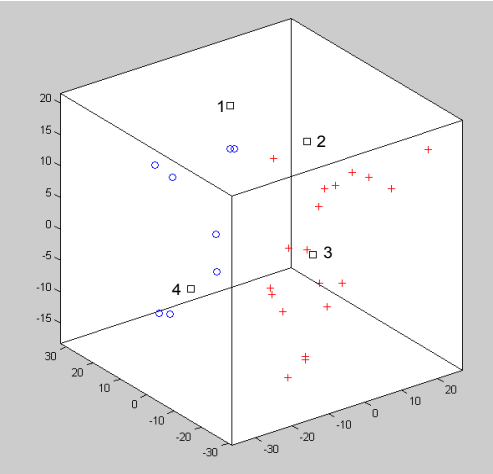


Figure 6.11: Principal component plot of the gene expression profiles obtained for the 31 melanoma tumours

by AP and the hierarchical algorithm. The concordance between the two methods appears satisfying.

Perou *et al.* breast cancer data

In the work of Perou and colleagues [137], an average-linkage hierarchical clustering method, as implemented in [47], was used to group the experimental samples on the basis of similarity in their patterns of expression. In the dendrogram derived from the hierarchical clustering, two large branches were apparent separating the tumour samples into those that were clinically described as ER positive and those that were ER negative. Within these large branches there were smaller branches for which common biological themes could be inferred, namely basal-like, Erb-B2+, normal-like and luminal epithelial/ER+ cancers.

As done in Lama *et al.* [107], missing values were imputed with the average values of neighbour genes resulted from a k Nearest Neighbours (k-NN) algorithm, setting $k = 10$ (R software package `EMV`, see [167]). The 10 genes 'more similar' to the one with missing value were selected on the basis of the sample correlation: a measure which conforms well to the intuitive biological meaning of 'co-expression'. In this way, all the unavailable data could be recovered and a 1753 x 65 matrix obtained, where rows represented genes and

		AP CLUSTERS	
PREVIOUS WORK'S CLUSTERS		1	2
	ER positive	26	22
	ER negative	8	4

Table 6.3: The distribution of subjects between new and old classification (Perou *et al.* data)

columns represented samples.

The AP algorithm was run using $1 - \text{Pearson correlation}$ as a similarity measure and the effect of the value of *preference* against the number of clusters is shown in Figure 6.12. By the inspection of plateaus, two, three, four and six may be considered as numbers of clusters determining stable solutions.

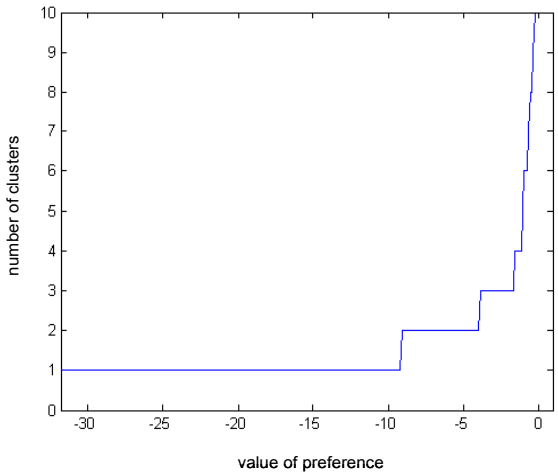


Figure 6.12: The effect of the value of input preference on the number of clusters for the Perou *et al.* data

When considering two clusters a clear correspondence with the ER positive and ER negative groups was not evident. In particular the distribution shown in Table 6.3 was obtained (please note that for 5 patients information about their ER status was missing).

Considering three clusters, the ER negative group was basically identified by the first cluster, which, on the other hand, contained also several ER positive patients (see Table 6.4 for details).

For four clusters a comparison with the four groups determined by the authors was done: from Table 6.5 it can be seen that all the basal-like tumours were captured by

		AP CLUSTERS		
PREVIOUS WORK'S CLUSTERS		1	2	3
	ER positive	8	17	23
	ER negative	10	2	0

Table 6.4: The distribution of subjects between new and old classification (Perou *et al.* data)

Cluster 1 and the normal-like tumours were assigned to Cluster 2. Clusters 3 and 4 seem to be characterised by the luminal/epithelial cancers. Instead, the Erb-B2 group was not captured by a single cluster by the AP. It is important to note that in their work, Perou and colleagues assigned three samples to the ER negative group without specifying which of the four subgroups they are part of.

		AP CLUSTERS			
PREVIOUS WORK'S CLUSTERS		1	2	3	4
	Basal	8	0	0	0
	Erb-B2	3	2	1	1
	Normal	0	11	0	0
	Luminal	7	3	16	10

Table 6.5: The distribution of subjects between new and old classification (Perou *et al.* data)

Finally, when Perou *et al.* dataset was divided in six clusters it was found what is reported in Table 6.6. Also with this classification in six groups, the basal-like patients were assigned to Cluster 1. Instead the Normal ones were divided in Clusters 2 and 5. Clusters 3, 4 and 6 seemed to be three Luminal groups. Moreover, the Erb-B2 group was not captured by a single cluster by the AP.

		AP CLUSTERS					
PREVIOUS WORK'S CLUSTERS		1	2	3	4	5	6
	Basal	7	0	0	0	1	0
	Erb-B2	2	2	0	1	0	2
	Normal	0	5	0	0	6	0
	Luminal	2	2	15	10	1	6

Table 6.6: The distribution of subjects between new and old classification (Perou *et al.* data)

		AP CLUSTERS	
PREVIOUS WORK'S CLUSTERS		1	2
	ER positive	64	11
	ER negative	5	37

Table 6.7: The distribution of subjects between new and old classification (van't Veer *et al.* data)

van't Veer *et al.* breast cancer data

In the work of van't Veer *et al.* [170], an unsupervised hierarchical clustering algorithm allow the authors to cluster 117 tumours on the basis of thousands of genes. For clustering, the 'one minus correlation' distance was used by the authors as described in the Supplementary Information for [170]. In this way two subgroups of breast cancer were detected, differing in ER status (ER-positive and ER-negative) and lymphocytic infiltration (see Figure 1a in [170]).

In this dataset, 293 genes had missing information for all 117 patients. These genes were excluded for the analysis. Other missing values were detected in the remaining data and, as done for the Perou *et al.* data, they were imputed with k-NN algorithm. In this way, the data matrix had 117 tumours and 24188 genes of log ratio of the intensities of the red and green channels.

Over van't Veer *et al.* dataset, the affinity propagation algorithm was applied in order to verify the two ER groups, obtaining the plot in Figure 6.13 showing the effect of the value of *preference* on the number of clusters. From this plot, looking at the plateaus, solutions with two, three, five, six and seven clusters were suggested. It is worth noting that AP was not able to indicate a solution with four clusters but one in six groups was evident in accordance with results previously obtained with the Perou *et al.* data.

When considering two clusters a very good concordance with the two groups obtained separating the ER positive and ER negative patients could be seen. In fact, just 16 over 117 cases did not show such a correspondence (see Table 6.7).

As reported in Table 6.8, the split in three clusters is nothing more than a subdivision of the previously determined cluster 1 into two subgroups.

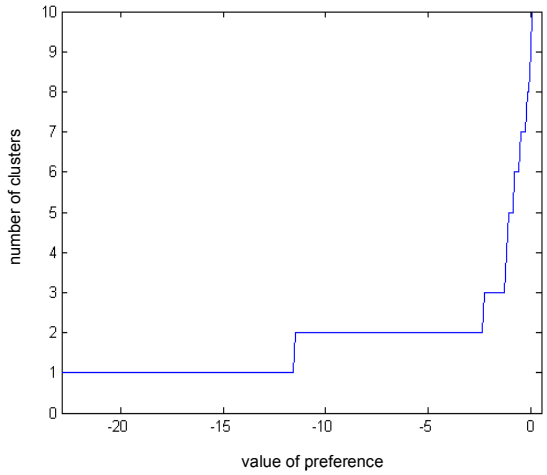


Figure 6.13: The effect of the value of input preference on the number of clusters for the van't Veer *et al.* data

		AP CLUSTERS		
PREVIOUS		1	2	3
WORK'S	ER positive	30	35	10
CLUSTERS	ER negative	1	4	37

Table 6.8: The distribution of subjects between new and old classification (van't Veer *et al.* data)

Nottingham breast cancer data

On the same data described in [1], where a hierarchical algorithm was used to classify patients in six different groups, the AP technique was applied and a plot of the effect of *preference* value on the number of cluster is reported in Figure 6.14. Looking at the plateaus, several solutions were proposed, dividing the dataset in two, three, four, five, six and eight clusters.

The solution with two clusters corresponded to the well known grouping of breast cancer in ER positive and negative tumours [137]. From the boxplots reported in Figure 6.15 this distinction is very well marked.

The solution with three clusters was very similar to the previous one, with a simple subdivision into two subgroups of the ER positive cluster.

The most interesting solution, in terms of comparison with the original work, was the one obtained using a preference value to get six clusters. The original groups determined

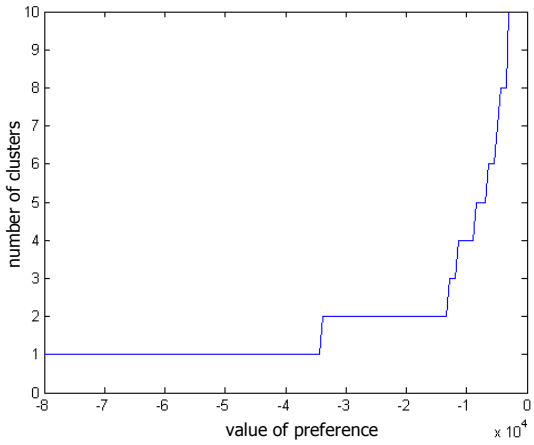


Figure 6.14: The effect of the value of input preference on the number of clusters fo the Abd El-Rehim *et al.* data

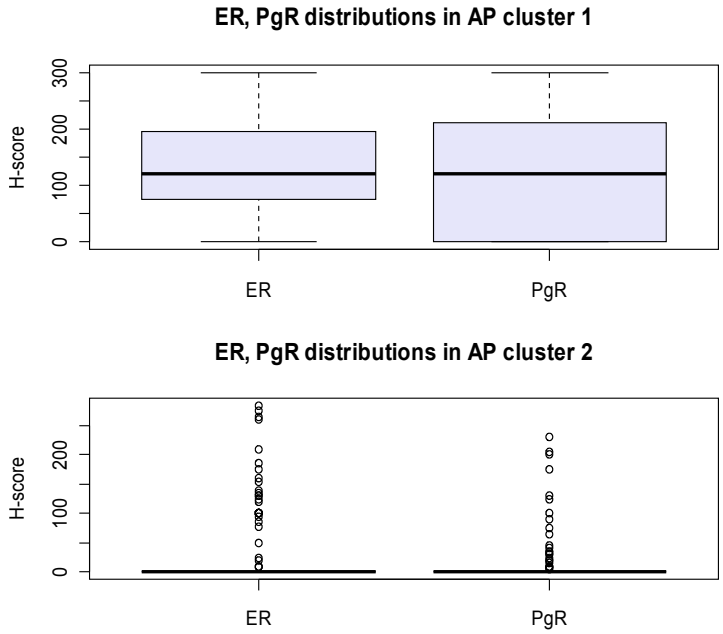


Figure 6.15: Boxplots of ER and PgR in AP clusters

in [1] can be described as two luminal and ER positive groups, two HER2 positive groups, which show differences in MUC1 and E-cadherin expressions, and a strong basal epithelial cluster. The last group, which contains only four patients, appears to be characterised by a basal phenotype with high p53. The distribution of patients assigned to the six clusters by AP and by the original hierarchical clustering is reported in Table 6.9. Labels to

the AP groups were assigned resorting to the weighted kappa index for the agreement between classifications [30] and the ordering for which the highest index was obtained was considered (weighted kappa = 0.302). Although there is not a good agreement between the two classifications, a couple of AP groups can also be described in terms of the original work. In particular, AP cluster 1 and 2 seem to express luminal characteristics, while AP group 5 appears to be a basal group.

		AP CLUSTERS					
		1	2	3	4	5	6
PREVIOUS WORK'S CLUSTERS	Luminal A	140	54	4	126	0	12
	Luminal B	58	113	3	5	1	0
	HER2 / MUC1-	38	10	40	24	5	22
	Basal 1	0	0	2	0	0	2
	Basal 2	2	22	59	0	94	6
	HER2 / MUC1+	28	78	31	27	2	68

Table 6.9: The distributions of subjects between new and old classifications (Abd El-Rehim *et al.* data)

On a more detailed analysis, the following results emerged. From Table 6.9, it is possible to see that also AP cluster 3 seems to express basal characteristics. Plotting the distributions of the main markers for patients assigned to groups 3 and 5 by AP, the Figure 6.16 is obtained. From this plot it is evident that the main difference between these two groups is the under/over expression of the p53 marker. These two groups seem to express characteristics which are typical of triple-negative patients. Considering only this subset of patients, it could be found that the 91% of them was assigned to either cluster 3 or cluster 5. These results once again confirm what it was recently found about p53 and triple negative patients (see Section 4.7 for details).

When using a preference value to get 4 AP groups, the distribution of patients was compared with the one obtained by the application of the PAM algorithm in the analysis described in Chapter 2 of this thesis. This comparison was chosen because, for the data analysed, PAM was the technique that produced the most stable and separated clusters, and for which there was an overall agreement between all the computed validity indices on the number of groups to consider. Results of this comparison are reported in Table 6.10. The overall agreement between the two classifications was once again calculated using

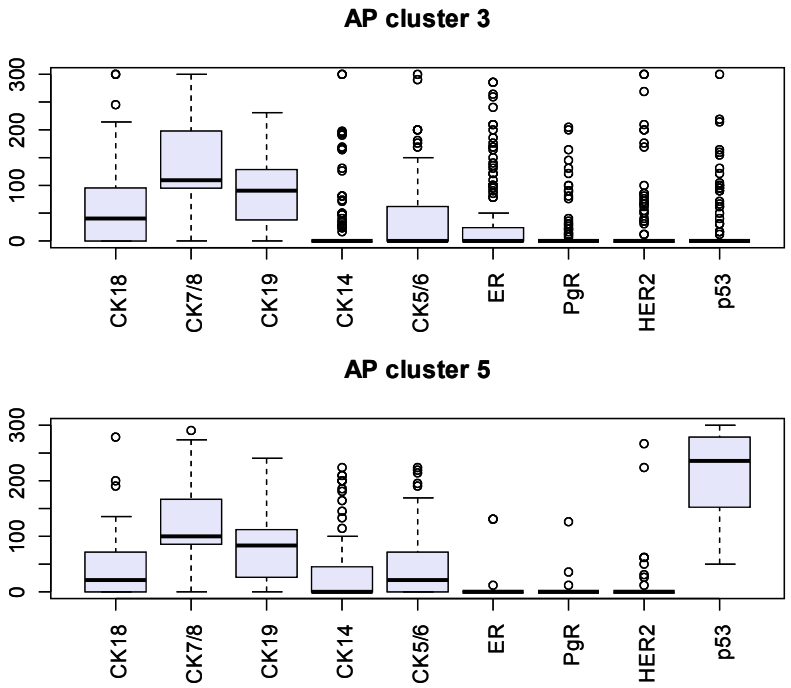


Figure 6.16: Boxplots of main markers for AP clusters 3 and 5

		AP CLUSTERS			
		1	2	3	4
PAM CLUSTERS	1	196	133	52	1
	2	37	254	24	9
	3	26	2	108	17
	4	30	8	58	121

Table 6.10: The distributions of subjects between new classification and PAM grouping (Abd El-Rehim *et al.* data)

the weighted kappa index, which in this case had a score of 0.56.

When using preference values for AP to obtain five and eight clusters, the groups that emerged were simply subdivisions of previously obtained groups.

6.3.3 Evaluation of CPU time

The affinity propagation algorithm is a novel approach also for what concerns the choice of the appropriate number of clusters. Rather than iterating the method using each time a different number of groups, the AP can be run iterating on the values of *preference*. However, these different approaches are also reflected on their time complexities. In

order to evaluate whether AP should or should not be included in the proposed framework for the identification of representative groups in a generic dataset (described in the next chapter), a comparison of the CPU time requested to perform both AP and K-means on the 'Nottingham dataset' was carried out.

To perform the experiments, the dataset of 25 biomarkers available for 1076 patients presented in [1] was used. The analysis consisted of several steps: at each of these, a bigger amount of data was considered, starting from 100 patients (randomly selected from the whole dataset) and increasing the size by adding 100 more patients (always selected at random from the remaining ones) at each step. The whole process terminated with the whole data being analysed. During each of these steps, data was clustered using K-means and AP. The former algorithm was run with the number of clusters varying from two to twenty, while AP using different *preference* values in order to get the same range of groups. The CPU time needed for each run was recorded and the average value then considered for the comparison. It is important to note that, the CPU time requested by AP was only compared with that of K-means because, as shown in Figure 6.17, the CPU times for K-means and PAM were quite similar. Results are shown in Figure 6.18, from which it can be seen how the CPU time needed for AP seems to follow an exponential behaviour, compared to a more linear one associated with the K-means. Based on this result, it was decided not to consider the AP algorithm in the proposed framework for elucidating core classes in a dataset.

The interesting outcome shown by Figure 6.18 is in contrast with the claim made by Frey and Dueck in [60], where the affinity propagation technique is described as feasible for large data sets and faster than hundreds of runs of the K-means method. The difference between the results shown in Figure 6.18 and those reported by Frey and Dueck in their original paper [60] may be due to the size of the data used not being big enough. However, the choice of not including the AP in the proposed framework was taken only considering its effective speed on the problem under investigation.

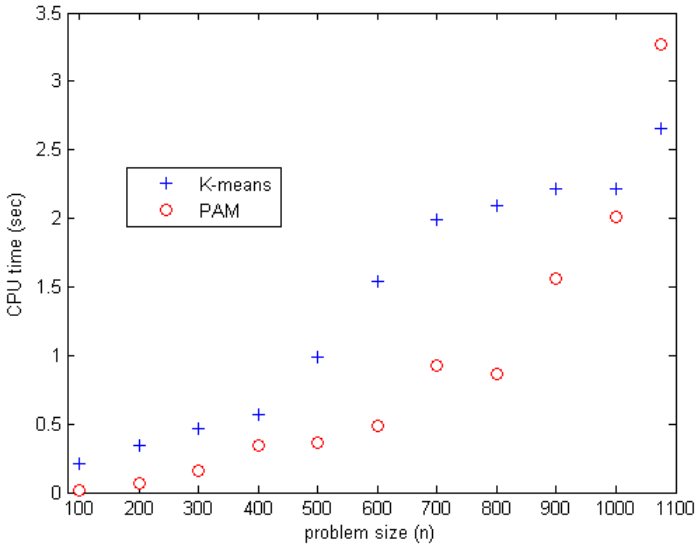


Figure 6.17: Comparison of CPU time between K-means and PAM

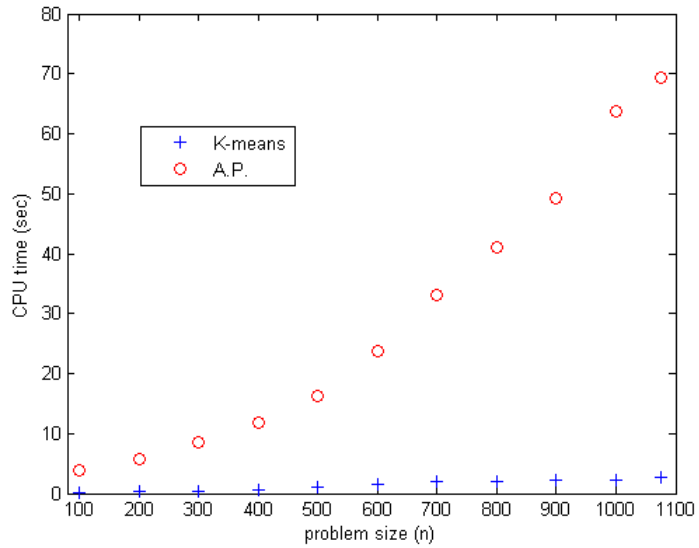


Figure 6.18: Comparison of CPU time between AP and K-means

6.3.4 Discussion of results

Cluster analysis is a powerful technique to explore complex diseases and improve prognosis. The recent literature on omic data is rich of new methods of cluster analysis able to deal with huge datasets. Moreover techniques of visualisations are usually adopted to suggest the number of clusters [47]. At the same time many papers warn against the

possible misuse of clustering techniques [70]. One of the main problems is the subjectivity of the analysis and the ability of clustering algorithms to create clusters even in absence of real structure. The choice of the number of clusters is one of the main problems to be faced when applying this kind of analysis. The possibility to use algorithms that incorporate a criterion for the choice of the optimal partition is one of the achievement of the recent developments in this research field.

The Affinity Propagation algorithm is characterised by a simple software implementation and it has the ability to suggest the cluster number. In addition, this algorithm has the advantage of taking into account real data points as exemplars; in this way, also categorical data may be analysed using, as similarity, a different distance from the Euclidean one. By looking for a set of exemplars, and not for a set of centroids, the search space is restricted, improving the computational efficiency of the Affinity Propagation.

The range of the suggested solutions reported in this chapter gives insights in the hierarchical structure of the data highlighting different levels of information for the treatment of cancer patients well in accordance with previous knowledge. For example, the solution with two clusters for Ferrara breast cancer data, evidenced in Figure 6.3, reflects the well known separation between tumours ER positive and negatives, as shown in Figure 6.4. This is a very important distinction and, in fact, in a number of papers of the pre-genomic era the number of clusters considered was in fact two [124, 140]. The solution with four clusters is in agreement with the solution selected in the previous work and the four clusters obtained are similar to that created by the PAM algorithm in [5]. The clustering obtained by AP on the melanoma data is able to reproduce the interesting findings of Bittner and colleagues having the advantage of avoiding any arbitrary choice due to the visual inspection of the dendrogram. Over Perou *et al.* the AP could almost reproduce previous classification: the basal-like tumours were captures by Cluster 1, and the normal-like ones were assigned to Cluster 2; Clusters 3 and 4 represent the luminal-epithelial tumours, while the Erb-B2 group was not identified by AP in a single cluster. The solution with two clusters for van't Veer *et al.* data reflects with a few number of exceptions, the separation

between ER+ and ER- groups evidenced in the original work. Finally, when considering Nottingham dataset, AP identified six groups as in the original work [1], and, in addition, reproduced the very important distinction within triple-negative patients. Two subgroups were identified, one over-expressing the p53 marker and the other being characterised by its absence.

The Affinity Propagation algorithm is claimed to be faster and more reliable than the K-means [60] when considering large data sets such as thousands of segments of DNA. For the problems under investigation in this study, however, this last assumption was not confirmed. As shown in Figure 6.18, the CPU time requested by the AP follows an exponential behaviour when increasing the problem size. K-means, instead, seems to have a linear trend.

6.4 Summary

In this chapter the Affinity Propagation clustering technique was described and its application over several known cancer data sets was presented to investigate the reliability of the algorithm in order to evaluate whether to include it in the proposed guideline for the elucidation of core classes. AP results were compared with those obtained with traditional algorithms. For two case studies the AP method suggested novel possible groupings which were not studied yet and that represent very interesting directions for future investigation. Moreover, it would be interesting to characterise the AP groups also in terms of clinical outcome and response to treatments. The computational complexity of the AP was also considered in this chapter and a comparison of the CPU times required for the AP and K-means computations over the 'Nottingham dataset' was reported. It was found that, for the problem under investigation, the AP algorithm required more CPU time. For this reason, this model-based technique was not included in the proposed framework for core classes elucidation.

The next chapter presents a step-by-step guideline to identify representing core classes

within any kind of data. By the application of different clustering techniques and using a consensus among the resulting groups a set of characteristic classes may be defined. These core patterns can be described using several statistical approaches and resorting to visualisation techniques and may be validated utilising different supervised algorithms.

Chapter 7

A Framework to Elucidate Core Classes in a Dataset

This chapter will provide a proposed algorithmic framework for elucidating a set of core groups in a general case study dataset. This proposed strategy will be explained in detail and validated over a novel set of breast cancer histone markers provided by the Division of Molecular and Cellular Sciences, Centre for Biomolecular Sciences, School of Pharmacy at the University of Nottingham.

7.1 Background and motivation

Clustering for real world problems and case studies has become a widely used approach to extrapolate important information from data and to separate different groups that share similar characteristics within them. Cluster analysis may be thought of as the discovery of distinct and non-overlapping sub-partitions within a larger population [130]. Different clustering techniques are known today, but, especially in breast cancer studies, researchers tend to focus on a single algorithm, usually the hierarchical one [137, 158, 159, 161, 170]. Choosing which method to use is not an easy task, as different clustering techniques return different groupings. When using more than one algorithm, it is then common to define a consensus across the results [98] in order to

integrate diverse sources of similarly clustered data [57].

At the same time, supervised classification techniques are widely used to learn a classification rule from a set of labeled cases (called the training set) to label new cases in a test set. Many different supervised classification methods have been developed in recent years, such as Neural Networks, Classification Trees, Bayesian Classifiers and many more.

Using machine learning algorithms and following similar approaches used in the past, a guideline to elucidate core classes in a general dataset was developed, in order to determine the fundamental characteristics of data expressed by different groups. At the beginning of this step-by-step guide, different clustering algorithms are applied and through a consensus clustering a set of common classes is defined. These core groups are then assessed using supervised classification techniques. In the next section this algorithmic framework is presented in detail and it will then be validated over a novel dataset.

7.2 Strategy

The proposed framework needs several input sets of methods and parameters, and it is formed by different steps which will be described below. The framework F has the following input arguments:

- The dataset under investigation Ω.
- The set of preliminary data analysis techniques and pre-processing algorithms P.
- The collection of several clustering techniques C which may be applied.
- The collection V of several validity indices which may be used to assess the grouping returned by cluster analysis.
- The set K of concordance measures (like kappa or rand indexes).
- The collection B of visualisation techniques to characterise the groupings.

- The set of several supervised learning techniques S.
- The statistical coefficient a to assess the association between groups and variables of interest.

Therefore, in its most general parameterisation, the framework may be written as

$$F(\Omega,\ P,\ C,\ V,\ K,\ B,\ S,\ a).$$

An organisation chart showing the overall approach and the logical steps used in this proposed pipeline is reported in Figure 7.1. Following this structure, each step of the framework is now presented.

1. In the first step, data preprocessing is performed. Rows which contain entries with missing values have to be deleted in order to run the clustering algorithms, and variables need to be 'homogeneous', which means that it is not convenient to have both numerical and categorical entries as part of the same variable distribution. If this happens, then clustering techniques may group together all numerical instances in one cluster and the categorical ones in another group, without emphasising other possible structures within the dataset. In this 'data preprocessing' step, several descriptive statistics (minimum, maximum, mean, median, quartiles, etc.) need to be checked as well, in order to have a complete picture of the data under investigation and to immediately spot any inconsistencies within them.

2. The second step is about clustering. Various unsupervised classification algorithms may be applied. In this work, four techniques are proposed to categorise cases into groups, namely the hierarchical (HCA), K-means (KM), Fuzzy C-means (FCM) and Partitioning Around Medoids (PAM). Given that K-means method is sensitive to cluster initialisation and in order to obtain reproducible results, this technique is initialised with the cluster assignments obtained by hierarchical clustering. All of the above techniques have been previously analysed and used (see Chapters 2 and

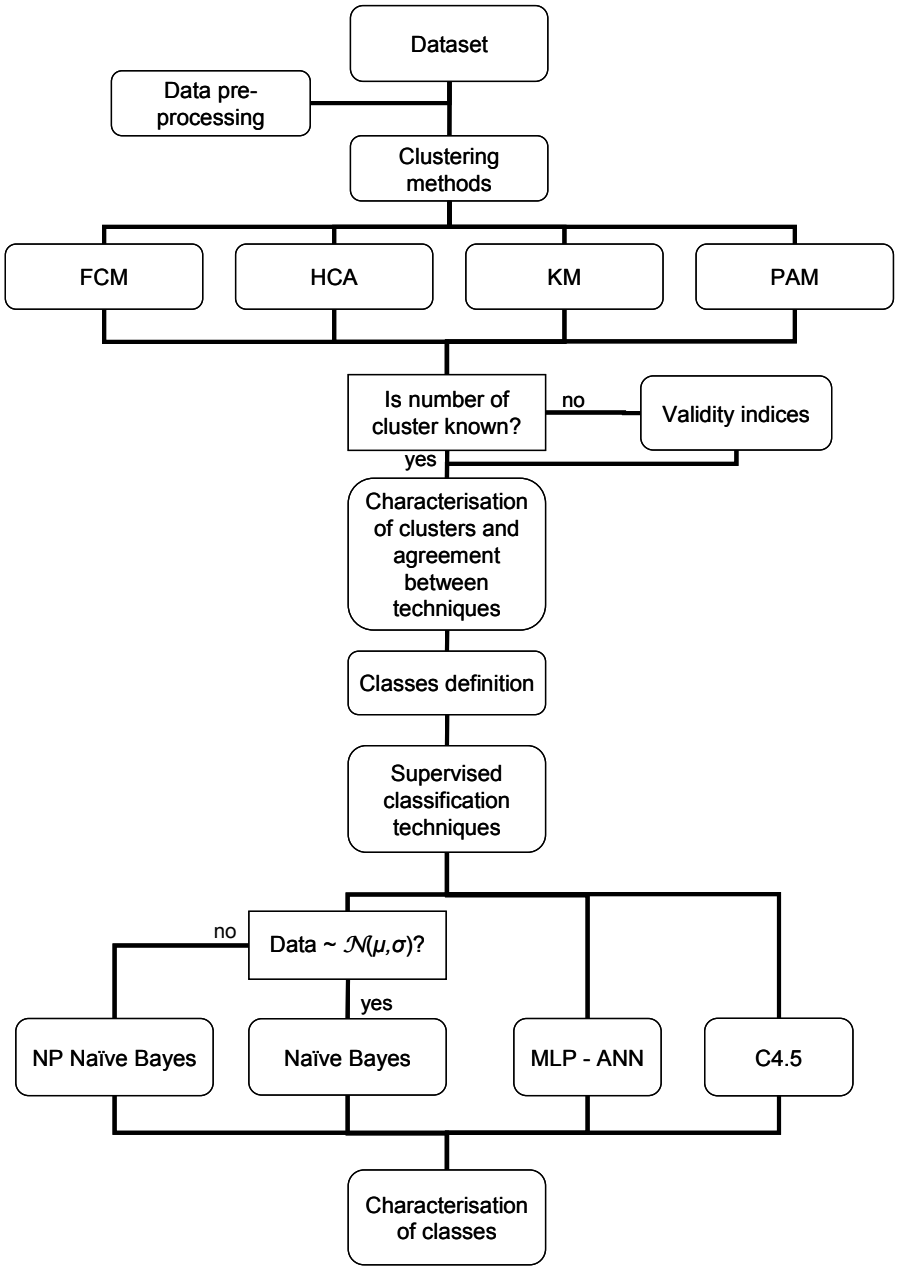


Figure 7.1: Organisation chart of the proposed framework

4 for details). This does not mean that these four are the best techniques to use, but they are among the most widely used clustering methods in machine learning and data mining. It is however a matter of fact that they performed quite well on the problems that were previously considered in this study.

3. In this step, validity indices are applied to clustering results. If the number of

clusters is not known before starting the analysis it is often convenient to resort to some external validation criteria. Several validity indices have been proposed in literature, but in this framework only six have been considered. The indices of Calinski and Harabasz [20], Hartigan [77], Scott and Symons [154], Marriot [118] and the two proposed by Friedman and Rubin [61] are used. According to specific rules, they indicate the appropriate number of groups to consider in the analysis. When indices indicate different numbers, it is possible to use them to rank in order the suggested groupings and then take the minimum sum of ranks as a form of agreement between indices.

4. When clusters are returned, a general characterisation of them can be obtained through visualisation techniques. Biplots, which are built considering the first two principal components and representing clusters projected on them, are a useful tool. This principal component (PC) technique is also used to reduce the dimensionality of a problem, as PCs account for as much of the variability in the data as possible. Another used technique for visualisation is the boxplot. It shows the distribution of each variable, computing its median value and lower and upper quartiles. Through the computation of the boxplots of all variables divided by clusters and using the biplots as well, it is possible to obtain a first 'informal' description of the grouping obtained by the clustering techniques. In addition, the agreement between the classifications obtained by different algorithms is, in this guideline, assessed either using the Cohen's kappa (κ) and weighted kappa (κ_w) indices [30, 31], or the Rand and adjusted Rand indices [85, 146], which also take into account the agreement occurred or corrected by chance. For the weighted-kappa index, weights are set in decreasing order from one (perfect agreement) to zero (complete disagreement) and all levels disagreement between raters are weighted according to their distance from perfect agreement. In any case, the length of the vector of weights must equal the number of rating categories. The adjusted Rand index corrects for the expected value of the Rand index of two random partitions not taking a constant value [194].

All these indices also give an indication about how likely will be to get a good consensus between classifications.

5. The following step reported in Figure 7.1 is related to classes definition. This is done via a consensus clustering which may be performed in several ways. In this proposed framework, the classifications obtained by different clustering algorithm are used and, looking at the biplots, the cluster labels are aligned in order to have the same patient assigned to the cluster named in the same way by different algorithms. Looking then at the same cluster number / label across all methods, core classes are defined by taking into consideration those cases assigned to the same group by different methods. These classes are aimed to include as many instances as possible.

6. To assess and verify the classes defined by the consensus clustering, several supervised classification techniques may be used. Among them, the C4.5 classifier (C4.5), the MultiLayer Perceptron Artificial Neural Network (MLP-ANN) and the naive Bayes classifier are considered in this framework. When data do not follow a normal distribution, the 'non-parametric' version of the naive Bayes (presented in Chapter 5) is used.

7. In the last step, the identified core classes are described resorting again to biplots and boxplots. When computing the biplots of classes, the 'not classified' cases usually are concentrated in the middle of the region. In addition, the correlation between classes and particular features of interest is computed resorting to the Phi (ϕ) statistics [53].

7.3 Validation over a novel dataset

To validate the approach presented in the previous section, the framework was applied in the following configuration: $(\Omega_1,\ P_1,\ C_1,\ V_1,\ K_1,\ B_1,\ S_1, \phi)$ where each input set is now described.

- Ω_1 = particular dataset provided by the School of Pharmacy at the University of Nottingham.
- $P_1 = \{$Missings deletion, descriptive statistics computation$\}$.
- $C_1 = \{$KM, PAM$\}$.
- $V_1 = \{$The same six validity indices as in Section 4.3.2$\}$.
- $K_1 = \{\kappa, \kappa_w\}$.
- $B_1 = \{$Biplots, boxplots$\}$.
- $S_1 = \{$C4.5$\}$.
- ϕ as the index to assess the association between classes and clinical variables available.

It is important to note that what follows is still an ongoing work, so not all of the techniques previously mentioned and reported in Figure 7.1 have been applied yet.

The dataset Ω_1 used to validate the proposed approach was a collection of 1254 consecutive breast tumours diagnosed from 1986 to 1998 included in the Nottingham Tenovus Primary Breast Carcinoma Series. Full details of the characterisation of the tissue microarray and the cohort of the patients are described elsewhere [50, 51]. Survival data were maintained on a prospective basis. Breast cancer specific survival was taken as the time (in months) from the date of the primary surgical treatment to the time of death from breast cancer [51]. Grading score was also available in this dataset. Breast cancer tissue microarrays were prepared as described in [2]. Each case was sampled twice from both the centre and the periphery of the tumour. Arrays of 150 cases per block were prepared. Breast cancer tissue microarray slides were prepared and immunohistochemically stained to detect the five histone markers as described in [1–3, 145]. The histone markers selected for this study were hMOF, SIRT1, H4K16ac, H3K9Me3 and SUV. They all have different functions: hMOF and SIRT1 are histone acetyltransferase and deacetylases in

orderly; they are responsible for H4K16ac acetylation. Later in this study, the SIRT1 marker will be dropped from the analysis. H4K16ac is a marker of active gene, while H3K9Me3 is a marker of silenced gene. Finally, SUV is the main factor responsible for H3K9 tri-methylation [113].

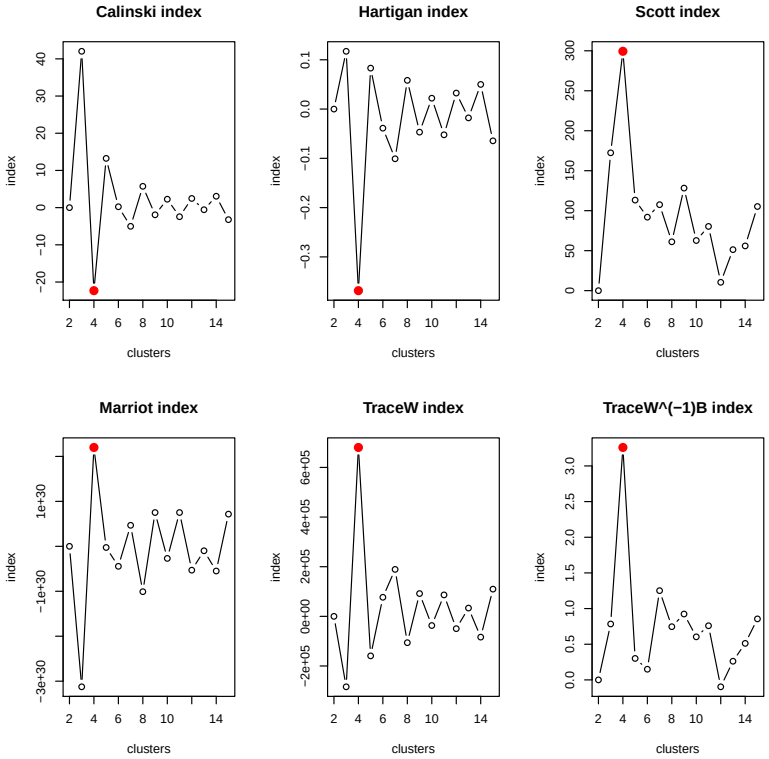
This collection of data presented many missing values; for the analysis described below, the five histone markers were only considered as well as those patients for which all the informations were present, thus reducing the number of patients to 301. The basic descriptive statistics like minimum, mean and maximum values for each feature were computed and together with the deletion of all missing values they formed the pre-processing techniques of the P_1 input set.

To assess the grouping, the K-means and PAM algorithms (see Sections 2.1.2 and 2.1.3) were applied with the number of clusters varying between two and twenty (the number of clusters is an explicit input parameter for both algorithms).

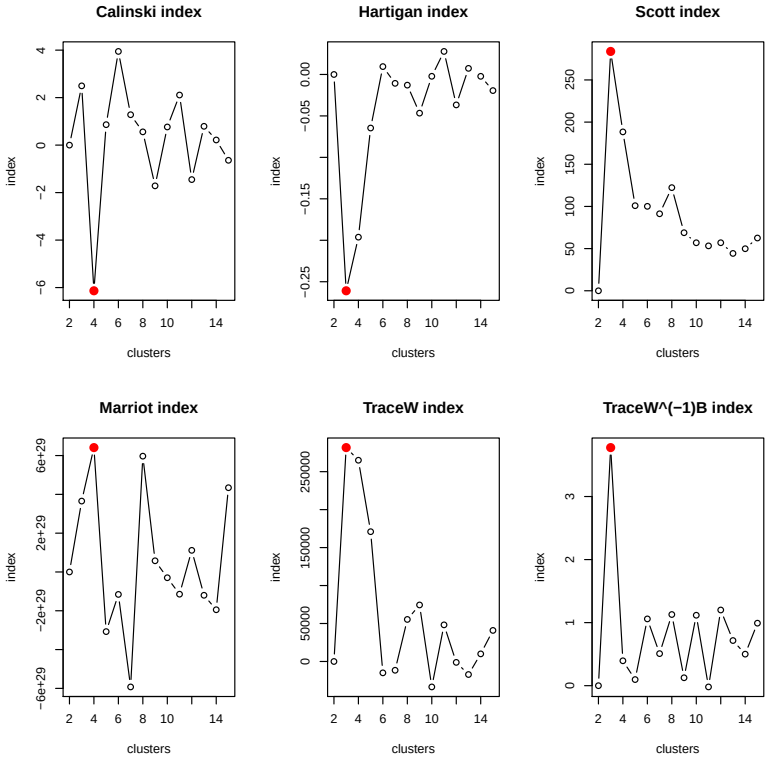
The same set of validity indices described in Section 4.3.2 was used for these experiments, as well as the same decision rules reported in Table 4.3 of the same section. The values of the indices for both K-means and PAM, for 2 to 20 clusters are shown in Figure 7.2; (a) shows the validity decision rule values obtained for K-means and (b) shows those obtained for PAM. The best number of clusters according to each validity index, for each clustering algorithm, is shown in Table 7.1 This corresponds to either the maximum or the minimum decision rule value (depending on the index), as indicated by the red point in Figure 7.2.

Index	K-means	PAM
Calinski and Harabasz	4	4
Hartigan	4	3
Scott and Symons	4	3
Marriot	4	4
TraceW	4	3
$\text{TraceW}^{-1}\text{B}$	4	3
Minimum sum of ranks	4	4

Table 7.1: Optimum number of clusters estimated by each index for K-means and PAM methods



(a) K-means indices behaviors

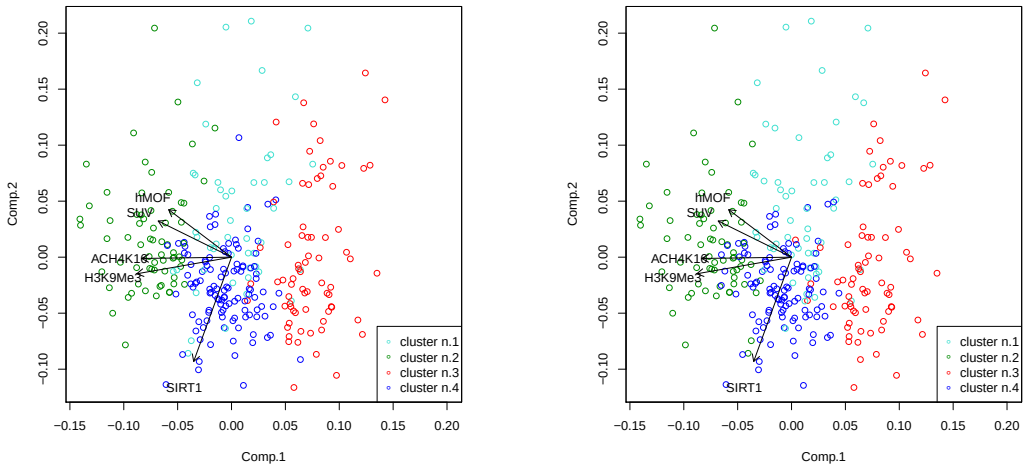


(b) PAM indices behaviors

Figure 7.2: Cluster validity indices obtained for K-means and PAM clustering, for varying cluster numbers from 2 to 20

From Table 7.1 it can be seen that all indices applied to the K-means results suggested to consider four groups, while such an agreement was not evident in the case of PAM algorithm. However, resorting once again to the minimum sum of ranks for the indices, it could be observed that both methods indicated four as the best number of clusters.

To visualise the results, biplots of the clusters were produced, where the same colours across different methods were used to minimise differences and to aid visualisation. Biplots for the solutions from each algorithm are reported in Figure 7.3. As reported in Section 4.4.3, the arrows in the plots represent the variables (markers) and their directions indicate in which group they are more expressed. From these plots, it can be clearly seen that for both algorithms cluster 3 seems to be characterised by low values of all the five markers, while cluster 2 appears to contain those patients with high values of covariates.

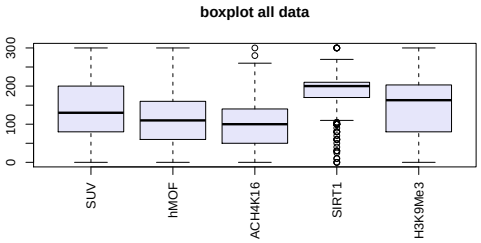


(a) K-means clustering (b) PAM clustering

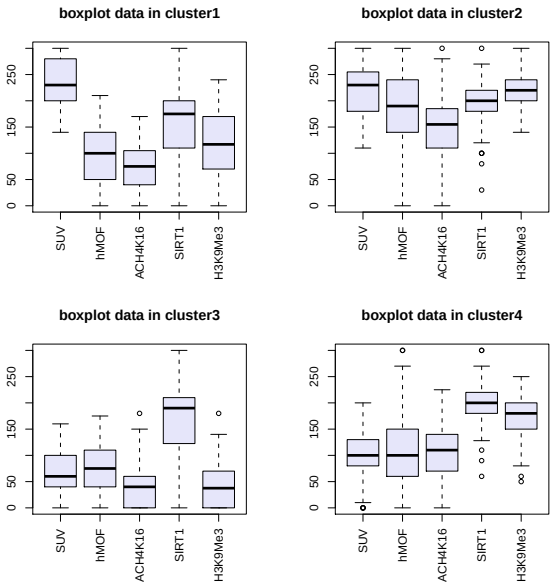
Figure 7.3: Biplots of clusters projected on the first and second principal component axes

To verify the last assumption, boxplots of the markers divided by cluster were produced and are shown in Figure 7.4. (a) shows the boxplot of the whole dataset, while (b) and (c) show, respectively, those obtained for K-means and PAM algorithms.

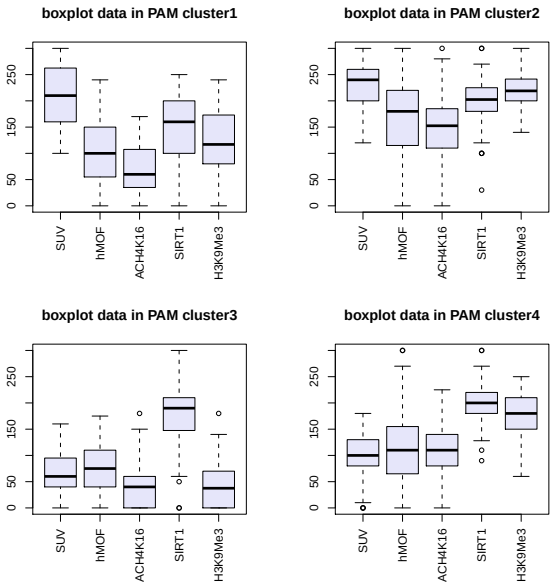
The cluster distributions (number of patients in each cluster) obtained for the K-means and PAM methods are shown in Table 7.2.



(a) Whole dataset



(b) K-means



(c) PAM

Figure 7.4: Boxplots for all markers, whole data and grouped by cluster for K-means and PAM methods

Cluster	K-means	PAM
1	49	55
2	67	68
3	72	72
4	113	106

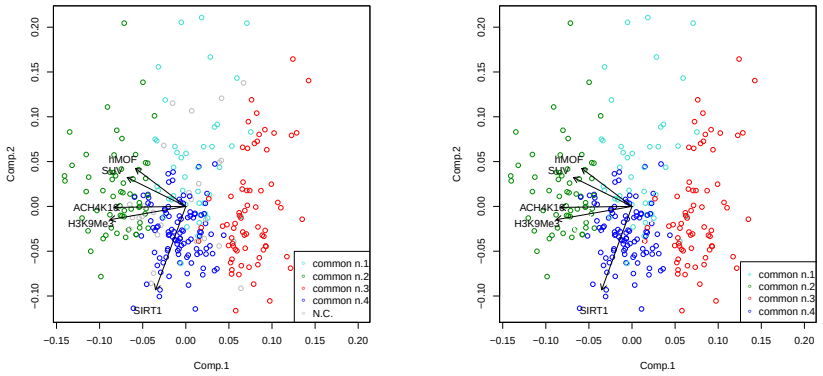
Table 7.2: Number of cases in each cluster

The correspondence of patients assigned in the four clusters solution for each of the methods was then examined. Cohen's kappa and weighted-kappa indices were computed to measure the degree of agreement among the two classifications derived by the two algorithms. Results for kappa and weighted kappa were, respectively, 0.882 and 0.870, showing a very good agreement between the two techniques used.

Focusing on the cluster correspondences, the aim was to define core classes containing the biggest possible number of patients. Considering the agreement among the clustering techniques and looking at those patients assigned to the same group by different methods, it was found that the sum of the number of patients assigned to the same group was 275 over 301 (91.4%). These results are again reflected in the kappa index values. The remaining patients (26, equal to the 8.6%) were placed into a 'not classified' (NC) group. As already pointed out for the analysis presented in Chapter 4, it must be stressed that the derivation of class assignments was made on the basis of the clustering results alone (which are, obviously, based on the five markers only). The distribution of patients in the four 'common' classes is reported in Table 7.3, together with the rule applied to define each class.

Class	No. of cases
1 (KM1 $\wedge$ PAM1)	44
2 (KM2 $\wedge$ PAM2)	61
3 (KM3 $\wedge$ PAM3)	69
4 (KM4 $\wedge$ PAM4)	101
Total number of cases assigned to classes 1 – 4	275
Total number of cases not classified	26

Table 7.3: Distribution of patients in the 'common' classes



(a) For all patients (b) For only patients in classes 1–4

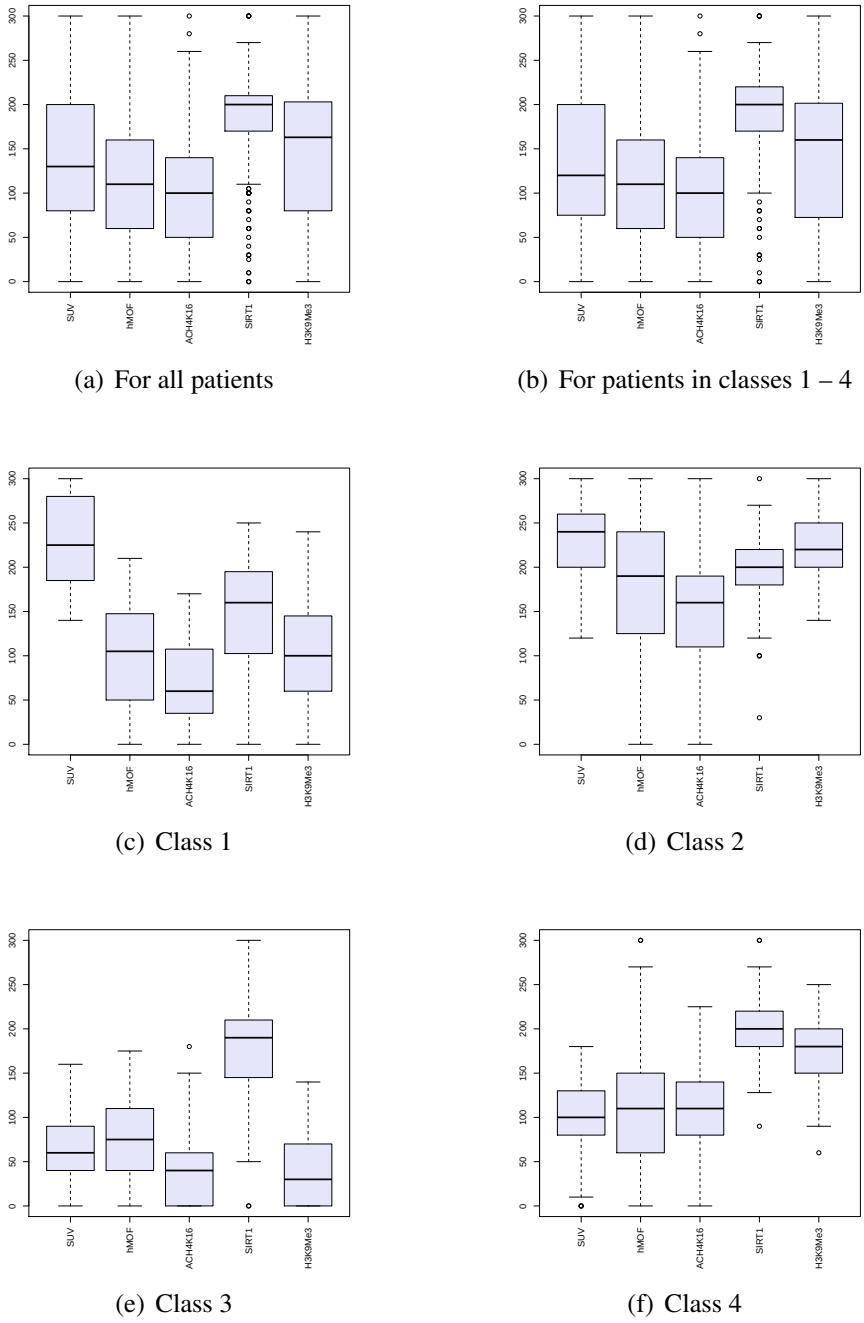
Figure 7.5: Biplots of classes projected on the first and second principal component axes

Biplots of the four consensus classes were produced and are reported in Figure 7.5, which provides a visualisation of the classes projected on the first two principal components.

Figure 7.5(a) shows the biplot obtained for all patients, in which the cases not assigned to any class (NC) have been coloured grey. It can be seen that these fall mainly into the centre region of the biplot. Figure 7.5(b) shows the biplot obtained for only patients assigned to classes 1 – 4. The first axis was mainly determined, on the left, by markers like ACK4H16 and H3K9Me3, while the second one is determined, on the bottom, by SIRT1 over-expression (although it can be seen from the boxplots that this marker is not characterising any particular group).

Figure 7.6 shows boxplots of all 5 markers, (a) for all cases, (b) for those cases assigned to classes 1 to 4, and (c-f) for each class separately.

By inspection of both the biplots and the boxplots, a ‘manual’ description of each class could be derived. First of all, it can be noticed that the distribution of SIRT1 remains almost identical throughout each class, thus suggesting that this marker is not playing any role in classes definition and characterisation. Then it appears evident that class 3 is mainly characterised by low expression of ACH4K16 and of H3K9Me3. Compared to the overall distribution of markers, class 2 seems to express high values for all the five covariates under investigation.



(a) For all patients

(b) For patients in classes 1 – 4

(c) Class 1

(d) Class 2

(e) Class 3

(f) Class 4

Figure 7.6: Boxplot for all markers, whole data and grouped by class

Starting from this consensus/common data, it was investigated whether it was possible to establish a set of rules to determine in which group a patient is more likely to be assigned starting from its variables values. To do so, the decision tree classifier C4.5 was used, in order to get some visual results too. This part of experiments was run using WEKA software [189]. The data set loaded was only formed by those patients that were

classified in one of the four groups previously defined, i.e. the 26 NC cases were not considered, thus leaving 275 instances to analyse. The C4.5 classifier was run ten times using the 10-fold cross validation option and the accuracy of the obtained classification was evaluated simply by looking at the percentage of the correctly classified instances and then averaging over the ten runs. The mean accuracy of C4.5 was 84.73%, such that 233 of the 275 patients were assigned to the proper class. A tree can be drawn from this classifier, with each node representing a variable and each leaf the class membership and the number of instances in it. The values reported on the connecting lines are the decision rules which are applied during the classification process. The tree generated by C4.5 algorithm is reported in Figure 7.7, where the minimum number of objects per leaf was left equal to 2 (default value).

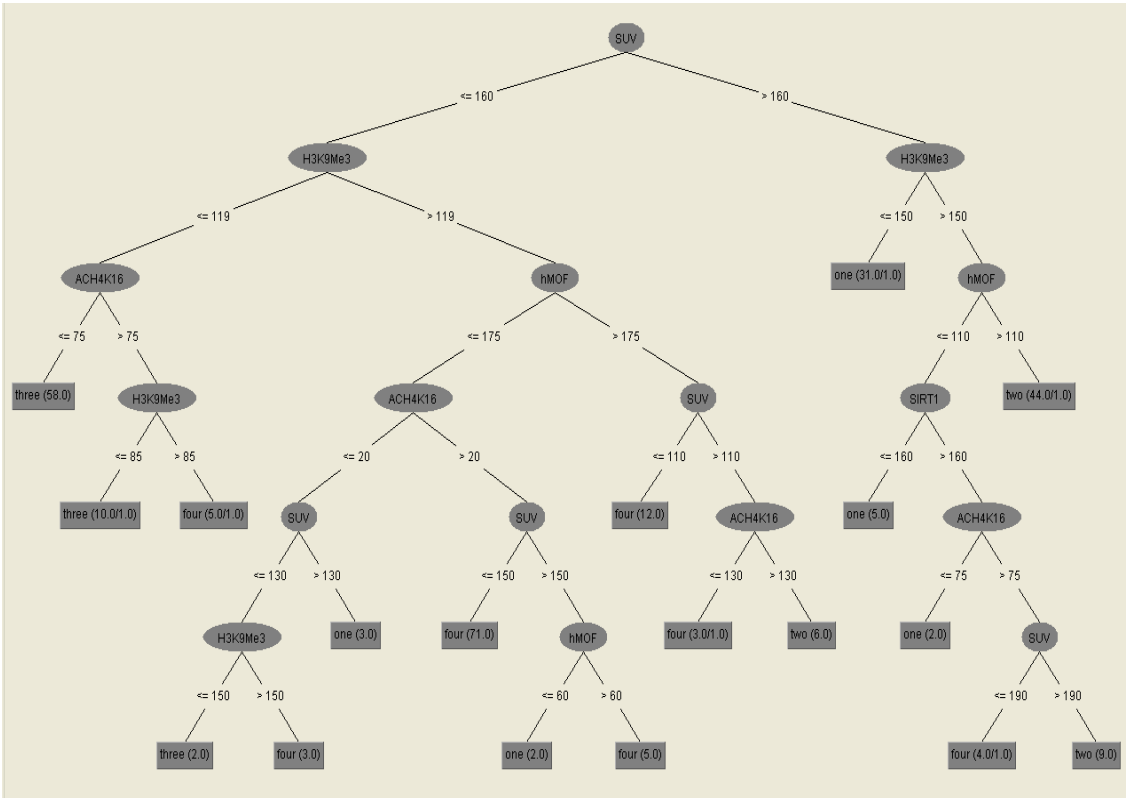


Figure 7.7: Decision tree generated by C4.5 for histone data

As it can be seen from Figure 7.7, the tree is very large and quite complicated to interpret. In such a situation it is possible to prune the tree, running again the algorithm and setting a higher number for the minimum number of objects per leaf. By doing so,

and increasing the minimum number to 8, the tree shown in Figure 7.8 was obtained.

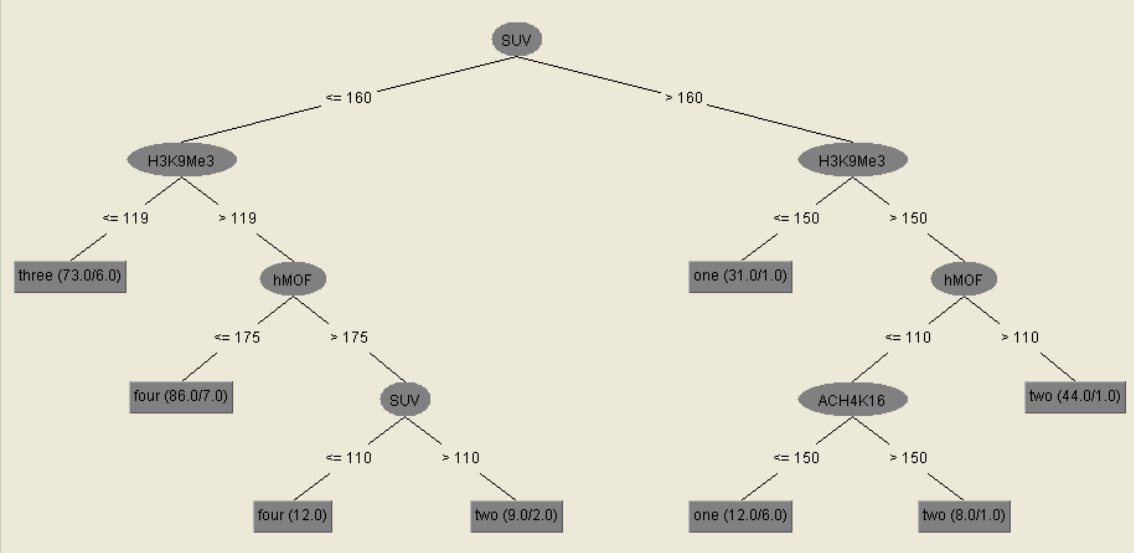


Figure 7.8: Decision tree generated by C4.5 for histone data (minimum number of objects per leaf = 8)

As previously noted, the SIRT1 marker does not play any relevant role in the classes definition. This is also confirmed by Figure 7.8, in which it can be seen that SIRT1 is not present as a node of the tree. For this reason, and following the clinical interpretation of this aspect, the SIRT1 marker was subsequently dropped from the analysis (see below for details).

As mentioned at the beginning of this section, several items of clinical information were also available for this study. In particular, the overall survival of patients was considered, and using the Kaplan-Meier estimator [94] the curves of the predicted survival against time were produced. The Kaplan-Meier estimator (also known as the product limit estimator) estimates the survival function from life-time data. In medical research, it might be used to measure the fraction of patients living for a certain amount of time after treatment [93]. A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant. For more technical details about Kaplan-Meier survival estimator see [93, 94].

The Kaplan-Meier curves obtained for this study are reported in Figure 7.9. It is important to note that several missing information about survival time and recurrence were deleted for the curves computation, leaving the total number of patients equal to 254.

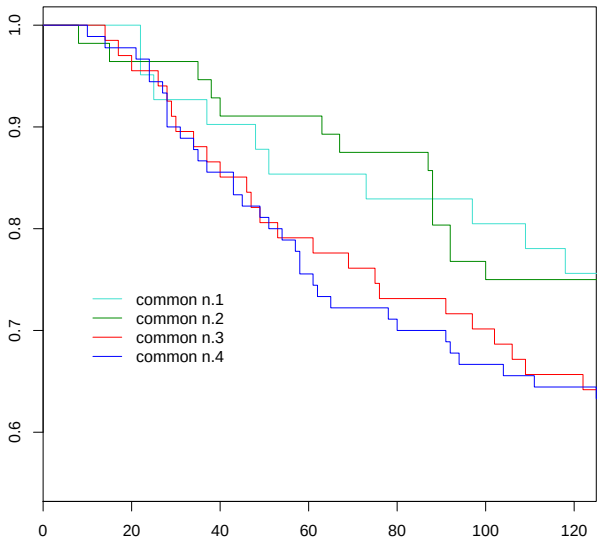


Figure 7.9: Kaplan - Meier curves for months of survival divided by class

The best overall survival is visible for patients grouped in classes 1 and 2, while classes 3 and 4 have the worst one. This somehow reflects what is shown in the boxplots (Figure 7.6), as class 2 presented high values for almost all the covariates, while class 3 had the lowest values.

As a last analysis, the association between tumour grade and classes was assessed, resorting to the Phi (ϕ) statistics [53]. Grading score is one of the components of the NPI score (Section 3.1.1 and [64]) and is determined by the Nottingham Grading System (NGS) which is based on the microscopic evaluation of tumour cells by pathologists [48, 144]. Grading is based on Mitotic Count, Tubule Formation and Nuclear Pleomorphism. Each of these indices can be scored between 1 and 3, therefore grading has its values in the range from 3 to 9. However, to simplify and to make grade comparable to other indices, the final values are scaled back from 1 to 3, according to rules reported in Table 7.4 [24].

Grade	*Combined Score*
Low Grade (1)	3 – 5
Intermediate Grade (2)	6 – 7
High Grade (3)	8 – 9

Table 7.4: Overall breast cancer grade by summation of all scores

The high grade patients are considered critical and their chances of survival are also poor [48].

In statistics, the phi coefficient ϕ is a measure of association for two binary variables. The phi coefficient is also related to the chi-square statistic for a 2×2 contingency table

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

where N is the total number of observations. The coefficient has a maximum value of one and the closer its value to one, the stronger the association between the two variables [53]. The association between grade and the common classes is reported in Table 7.5 (the total number of patients here is 273, as there were 2 missing information for Grade).

	Common classes				
	1	2	3	4	ϕ
Low Grade	8	19	7	26	
Interm. Grade	11	30	21	20	0.343
High Grade	25	12	40	54	

Table 7.5: Common classes distribution in relation to grading score

This table proves what is know from literature. As a matter of fact, it can be seen that the majority of high grade patients, which are known to have a poor prognosis [48], are grouped in classes 3 and 4, which, according to the Kaplan-Meier curves reported in Figure 7.9 are the groups with the worst overall survival.

As previously reported, SIRT1 did not play any role in the classification decision making. In accordance with researchers in the School of Pharmacy, after presenting them the above results, it was agreed to reconsider the whole problem without taking into account

the SIRT1 marker. In this way, several missing information could also be retrieved, having now a dataset with 347 patients, 46 more than previously.

The same steps as before were followed to perform the analysis, starting with the application of unsupervised clustering techniques and the definition of a consensus clustering to end with some correlation analysis. Running the K-means and PAM algorithms over this new dataset and computing the same set of validity indices as before, the plots shown in Figure 7.10 were obtained.

As it can be seen, all indices for PAM algorithm indicate three as the best number of clusters, while there is not such an agreement for the K-means method. The suggested number of clusters for each index in each method is reported in Table 7.6.

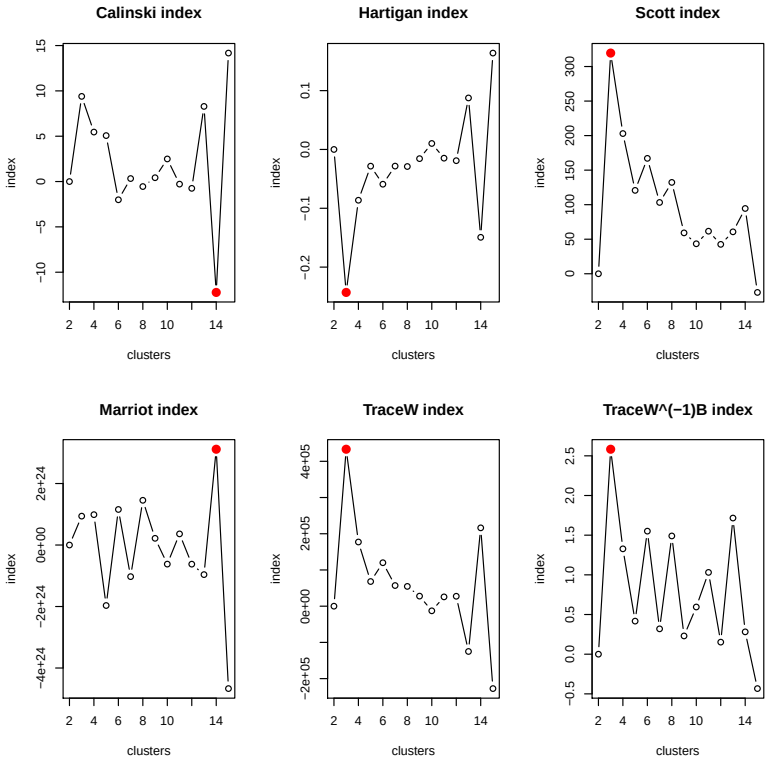
Index	K-means	PAM
Calinski and Harabasz	14	3
Hartigan	3	3
Scott and Symons	3	3
Marriot	14	3
TraceW	3	3
Trace$W^{-1}B$	3	3
Minimum sum of ranks	3	3

Table 7.6: Optimum number of clusters estimated by each index for K-means and PAM methods

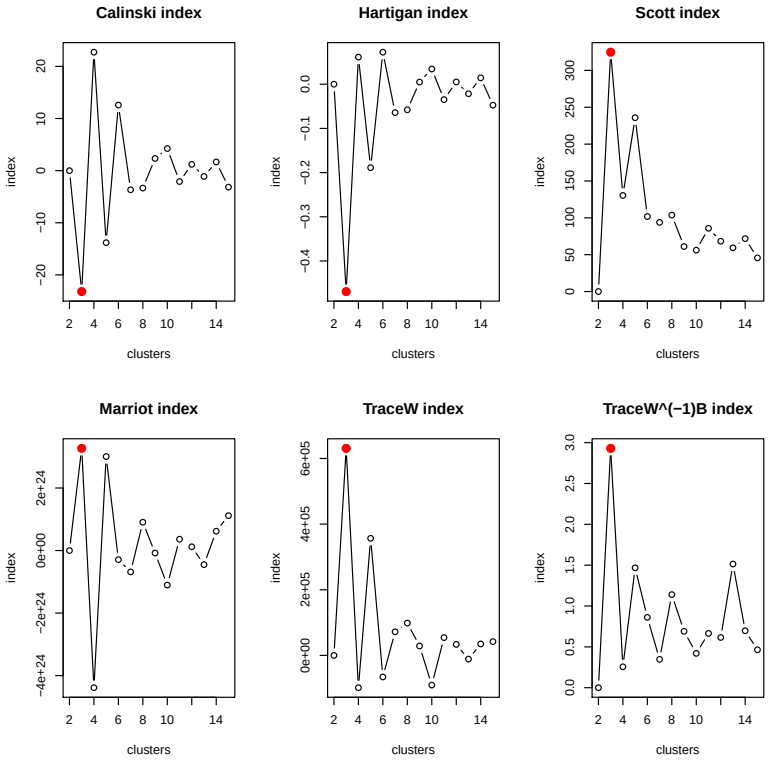
Resorting once again to the minimum sum of ranks for the indices, it can be observed that both methods indicated three as the best number of clusters.

For the visualisation of the results, biplots of the clusters were produced and are reported in Figure 7.11. From these plots, by looking at the overlap between arrows and circles, it can be seen that for both algorithms cluster 3 (in red) seems to be characterised by low values of all the four markers, while cluster 2 appears to contain those patients with high values of covariates.

To verify the last assumption, boxplots of the markers divided by cluster were produced and are shown in Figure 7.12. (a) shows the boxplot of the K-means groups, while (b) shows those obtained for the PAM algorithm.

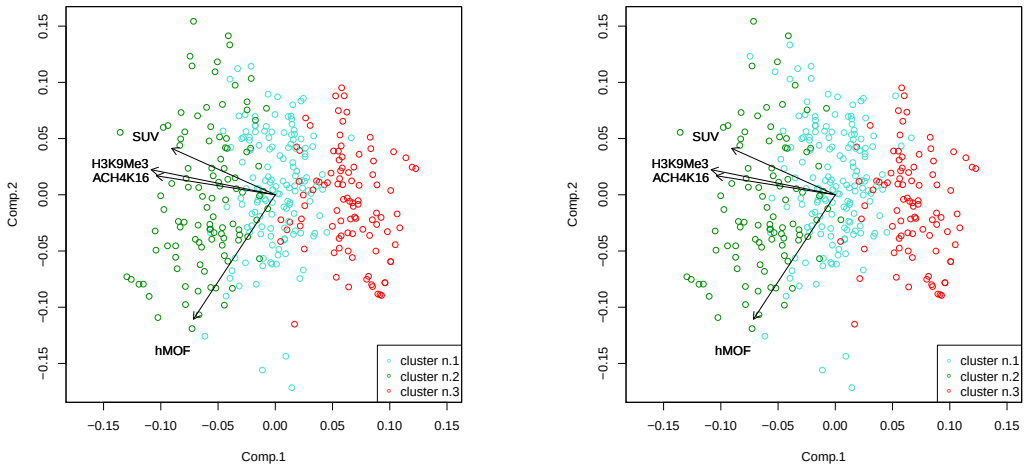


(a) K-means indices behaviors



(b) PAM indices behaviors

Figure 7.10: Cluster validity indices obtained for K-means and PAM clustering, for varying cluster numbers from 2 to 20



(a) K-means clustering

(b) PAM clustering

Figure 7.11: Biplots of clusters projected on the first and second principal component axes

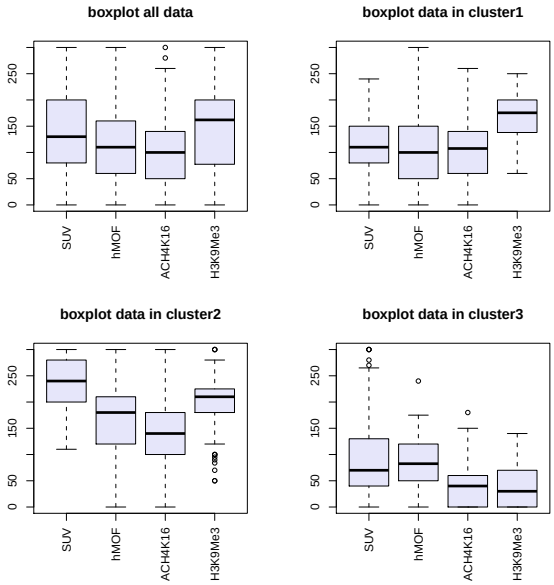
The cluster distribution (number of patients in each cluster) obtained for the K-means and PAM methods is reported in Table 7.7.

Cluster	K-means	PAM
1	144	161
2	105	96
3	98	90

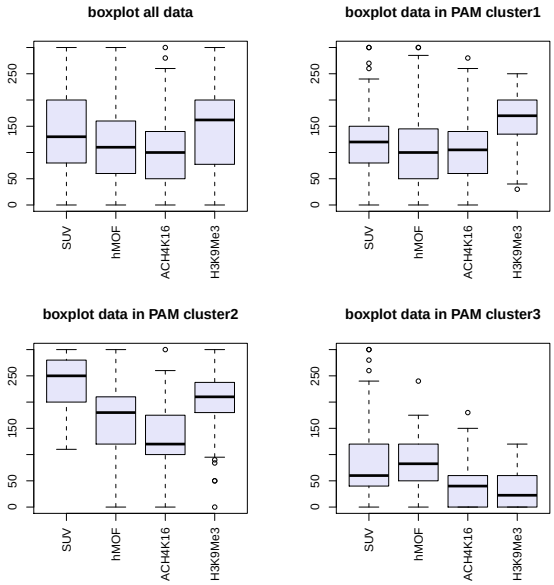
Table 7.7: Number of cases in each cluster

The correspondence of patients assigned in the three clusters solution for each of the methods was then examined resorting to both the unweighted and weighted kappa index κ. Results were, respectively, 0.911 and 0.906, showing almost a perfect agreement between the two techniques used.

Focusing on the cluster correspondences, core classes containing the biggest possible number of patients were defined. Considering the agreement among the clustering techniques and looking at those patients assigned to the same group by the different methods, three common classes were found containing the 94.2% of the overall population. In practice, 20 patients were not assigned to any of these three classes and were placed into a ‘not classified’ (NC) group. The distribution of patients in the three ‘common’ classes



(a) K-means



(b) PAM

Figure 7.12: Boxplots for all markers grouped by cluster for K-means and PAM methods

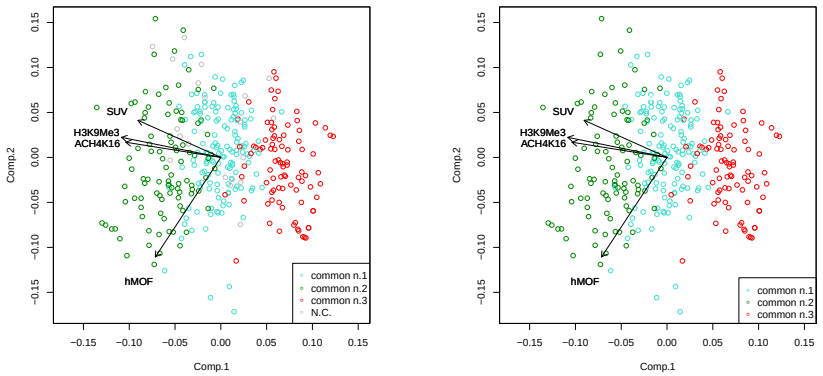
is reported in Table 7.8, together with the rule applied to define each class.

Biplots of the three consensus classes were produced and are reported in Figure 7.13. Again, Figure 7.13(a) shows the biplot obtained for all patients, in which the cases not assigned to any class (NC) have been coloured grey. It can be seen that these fall mainly into the centre-top region of the biplot. Figure 7.13(b) shows the biplot obtained for only patients assigned to classes 1 – 3. The first axis was mainly determined, on the left, by

Class	No. of cases
1 (KM1 ∧ PAM1)	143
2 (KM2 ∧ PAM2)	95
3 (KM3 ∧ PAM3)	89
Total number of cases assigned to classes 1 – 3	327
Total number of cases not classified	20

Table 7.8: Distribution of patients in the 'common' classes

ACK4H16 and H3K9Me3 markers, while the second one is determined, on the bottom, by hMOF over-expression.



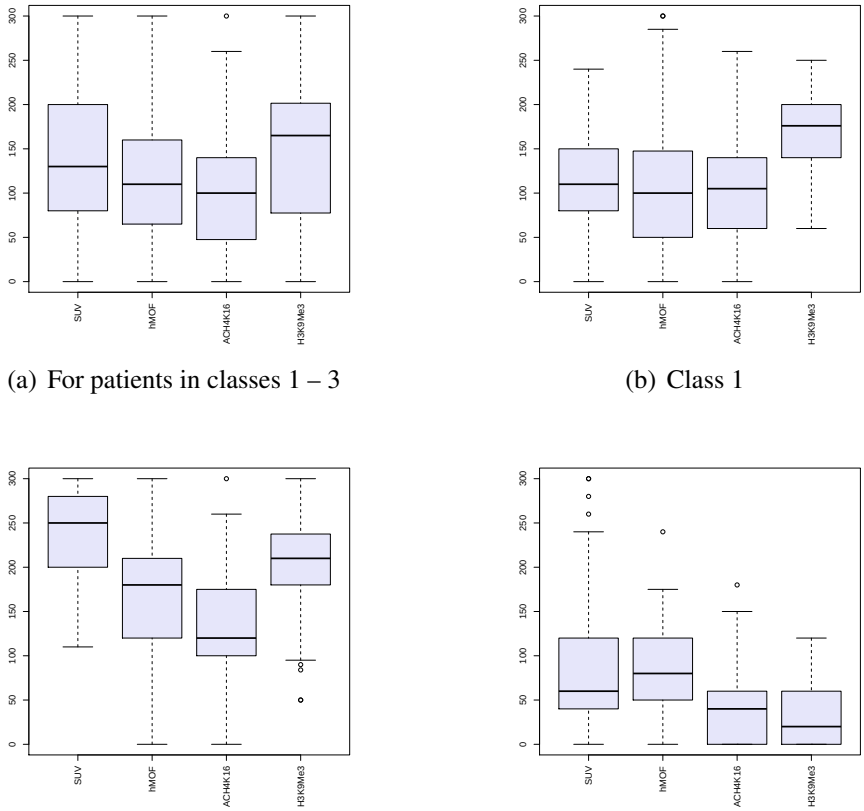
(a) For all patients (b) For only patients in classes 1 – 3

Figure 7.13: Biplots of classes projected on the first and second principal component axes

Figure 7.14 shows boxplots of all four markers, (a) for those cases assigned to classes 1 to 3, and (b-d) for each class separately.

By visual inspection of both the biplots and the boxplots, a 'manual' description of each class could be derived. It seems quite evident that class 3 is mainly characterised by low expression of all the four markers. Compared to the overall distribution, class 2 appears to express higher values while class 1 is quite similar, especially with respect to hMOF and ACH4K16.

Starting from this consensus/common data, it was again investigated whether it was possible to establish a set of rules to determine in which group a patient is more likely to be assigned starting from its variables values. To do so, the decision tree classifier C4.5 was used, running this algorithm in WEKA. The data set loaded was only formed by



(a) For patients in classes 1 – 3 (b) Class 1

(c) Class 2 (d) Class 3

Figure 7.14: Boxplot for all markers grouped by class

those patients that were classified in one of the three groups defined above, i.e. the 20 NC cases were not considered, thus leaving 327 instances to analyse. The C4.5 classifier was run ten times using the 10-fold cross validation option and the accuracy of the obtained classification was again evaluated using the percentage of the correctly classified instances and averaging over the ten runs. The mean accuracy of C4.5 was 90.8%, which means that 297 of the 327 patients were assigned to the proper class by this classifier. The tree generated by C4.5 algorithm is reported in Figure 7.15, where the minimum number of objects per leaf was left equal to 2 (default value). As it can be seen from Figure 7.15, also this tree is very large and quite complicated to interpret. So it was decided to prune the tree, running again the algorithm and setting a higher number for the minimum number of objects per leaf. By doing so, and increasing the minimum number to 8, the tree shown in Figure 7.16 was obtained. To obtain a smaller tree and simplify the rules for

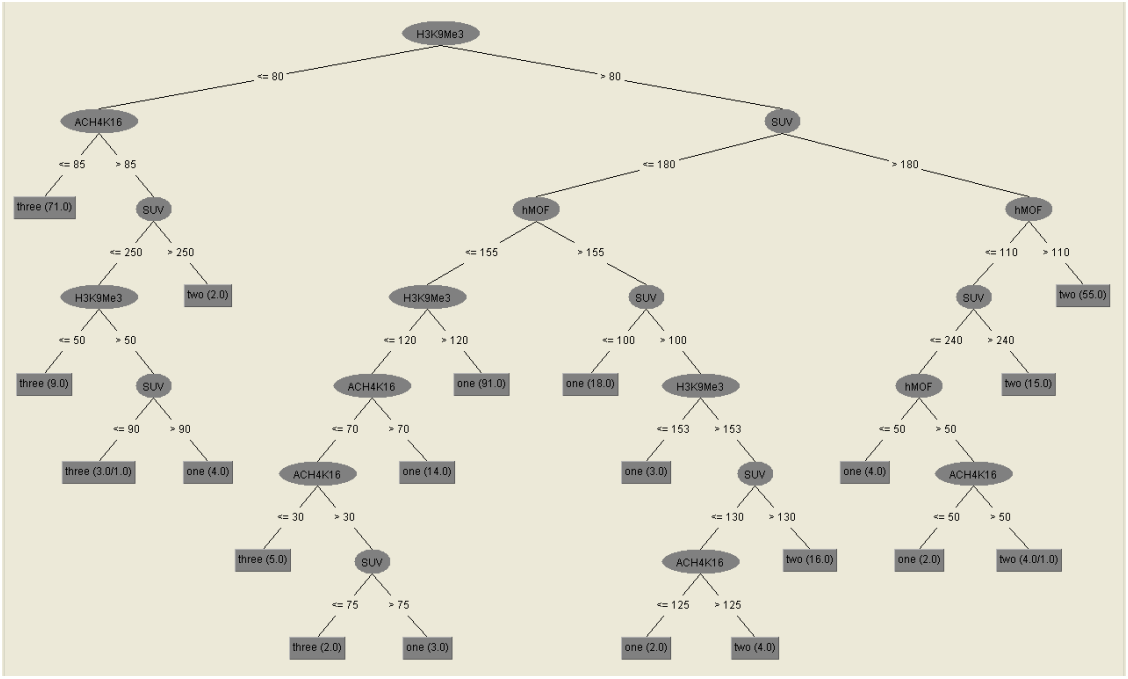


Figure 7.15: Decision tree generated by C4.5 for histone data

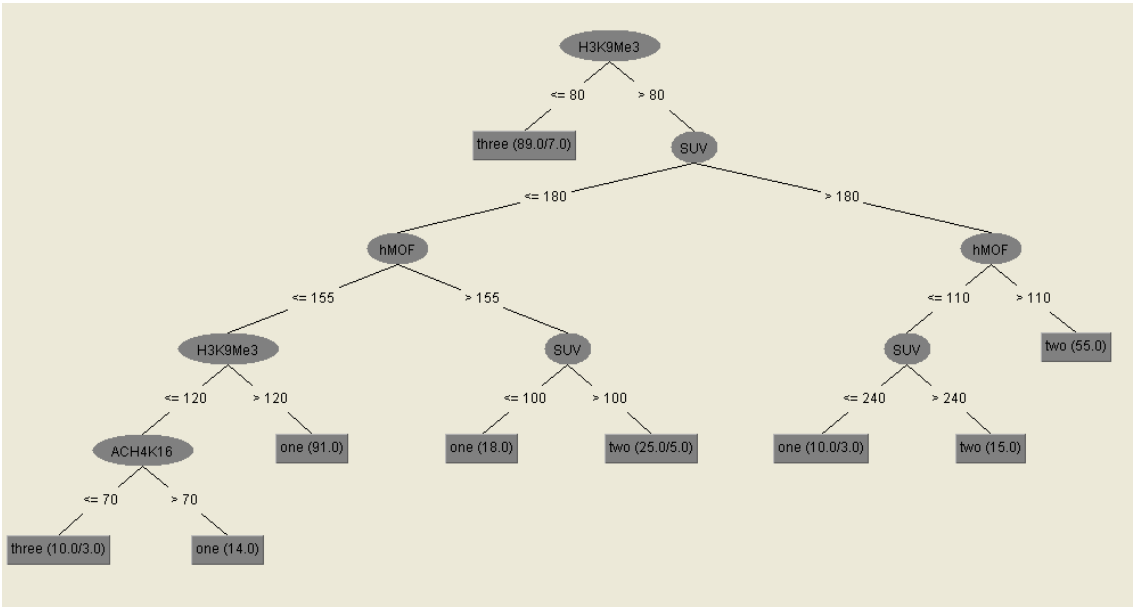


Figure 7.16: Decision tree generated by C4.5 for histone data (minimum number of objects per leaf = 8)

classification, a third tree was computed by running the classifier with 16 as the minimum number of objects per leaf. The result is visible in Figure 7.17. It is important to realise that, by pruning the tree, the accuracy of the classifier decreases, thus leaving one to decide whether a simple set of rules or a high accuracy is preferable for the analysis.

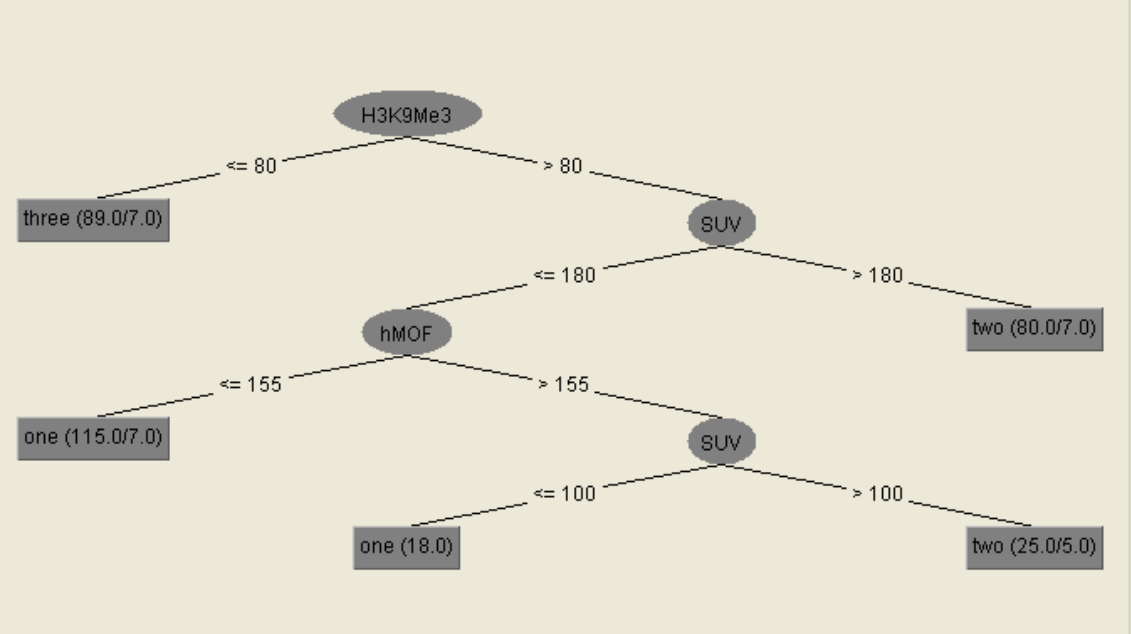


Figure 7.17: Decision tree generated by C4.5 for histone data (minimum number of objects per leaf = 16)

The same clinical information as before were available for this study. Again, the attention was focused on the overall survival of patients, and using the Kaplan-Meier estimator [94] the curves of the predicted survival against time were computed. These curves are reported in Figure 7.18. It is important to note that several missing information about survival time and recurrence were deleted for the curves computation, leaving the total number of patients equal to 319. The best overall survival is visible for patients grouped in class 2, while classes 1 and 3 have the worst similar survival. Once again, the class with the highest values of covariates (class 2) appears to have the overall best survival.

As a last analysis, the association between tumour grade and classes was assessed, resorting to the Phi (ϕ) statistics [53]. The association between grade and the common classes is reported in Table 7.9 (the total number of patients here is 324, as there were 3 missing information for Grade). This table proves once again what is know from literature. As a matter of fact, it can be seen that the majority of high grade patients, which are known to have a poor prognosis [48], are grouped in classes 1, which, according to the Kaplan-Meier curves reported in Figure 7.18 is the group with the worst overall survival.

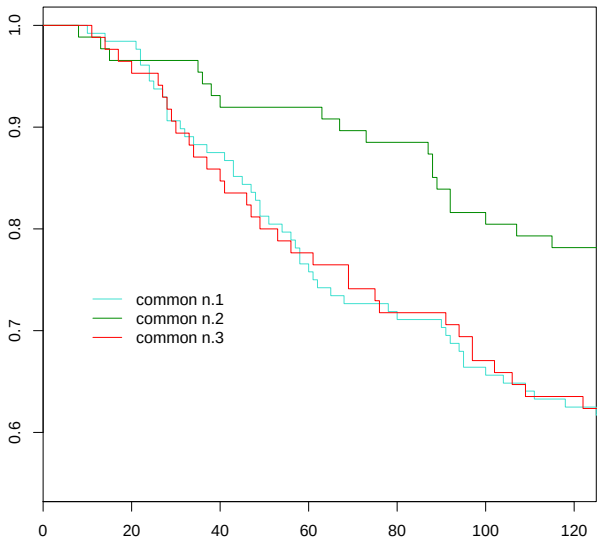


Figure 7.18: Kaplan - Meier curves for months of survival divided by class

	Common classes			
	1	2	3	ϕ
Low Grade	33	31	11	
Interm. Grade	35	42	30	0.292
High Grade	73	22	47	

Table 7.9: Common classes distribution in relation to grading score

Comparing the results obtained not considering the SIRT1 biomarker with those obtained at the beginning of this study, it can be seen that the three common classes identified by the consensus clustering on four markers dataset have a clearer definition. In fact the three groups are somehow characterised by low / intermediate / high markers levels. Moreover, the agreement between classifications (kappa and weighted kappa indexes) is higher for the four markers data than for the one with five biomarkers. This in part justifies the choice of removing SIRT1 from the analysis.

7.4 Summary

In this chapter a proposed framework for emphasize core classes within a dataset was presented. It follows a logical scheme in which at the beginning unsupervised clustering techniques are used to group patients (or any kind of data) in clusters which share similar characteristics. It is important to use more than a single clustering method, as it has been proved in literature [5] that different algorithms return different groups and it is not possible to say which, if it really exists, is the best clustering technique. By a visual inspection of the results and by using an index to assess the degree of agreement between different classifications, an informal consensus clustering may be derived, considering those patients assigned to the same group by different algorithms as 'in-class' and labelling all the others as 'not classified'. The resulting classes may then be analysed in different ways, either using again biplots and boxplots, or looking at their relations with other variables (which, in this study, were several clinical information). In any of those cases, an automated supervised classification technique may be used to confirm and assess the identified grouping.

In the second part of the chapter, the proposed approach was validated over a dataset of histone biomarkers which was provided by the Division of Molecular and Cellular Sciences in the School of Pharmacy of the University of Nottingham. It is important to note that this work is still in progress and the results obtained so far still need an accurate interpretation from clinicians and researchers at the Schools of Pharmacy and Molecular Medical Sciences. However, this study served to present and validate the proposed procedure, which, so far, has given very promising and encouraging results.

Chapter 8

Conclusions

This thesis focused on the development of advanced computational techniques for the classification of breast cancer into sub-types of the disease based on protein expression levels of selected markers. This multi-disciplinary research work aimed to identify breast cancer profiles with novel clinical relevance and to propose a new computational framework to elucidate core representative classes in a general dataset. To reach these objectives, at the beginning of this study different clustering algorithms were applied over a breast cancer case series to assess the stability of the resulting classification across different methods.

Clustering is the process of grouping a set of unlabelled multidimensional patterns (objects or data points), such that patterns in the same cluster have the most similar characteristics, and patterns within different clusters have the most dissimilar characteristics. Cluster analysis is a powerful technique to explore complex diseases and improve prognosis. The recent literature on 'omic' data is rich of new methods of cluster analysis able to deal with huge datasets. Moreover techniques of visualisations are usually adopted to suggest the number of clusters [47]. At the same time many papers warn against the possible misuse of clustering techniques [70]. One of the main problems is the subjectivity of the analysis and the ability of clustering algorithms to create clusters even in absence of real structure. However, in many studies, clustering techniques have been successfully

used to emphasise several breast cancer profiles [1, 5, 137, 158, 159, 161, 170].

Using clustering algorithms and a consensus between the resulting classifications, six diverse groups were identified and their characterisation appeared to be somehow novel. Two of these classes, in fact, were not emphasised in literature yet. After considering several supervised learning methods and developing a new algorithm to cope with non-normality of the data in many real world problems, a framework to highlight and validate representative core clusters within any kind of data was developed. This proposed guideline, together with its application over a novel case study, has been presented in this thesis.

In the next section, the main contributions of this work are reported, followed by several directions for future research that may be followed to complement this thesis. The chapter is concluded with the dissemination resulted from this research work.

8.1 Contributions

This work has resulted in the following contributions.

- **Comparison of several clustering techniques for breast cancer data**.
  Hierarchical clustering, K-means, Partitioning Around Medoids, Fuzzy c-means and Adaptive Resonance Theory were analysed and applied over a novel breast cancer data set. It does not seem that a systematic comparison of these techniques on breast cancer has been performed in literature yet. In many works [1, 47, 137, 158, 159, 161, 170], hierarchical algorithm has been used to detect and characterise breast cancer phenotypes using both gene expression profiles and tissue microarray approach. One limitation of this technique is its subjectivity in the clusters identification, since there is not a standard criterion or algorithm for choosing a cutoff point for a dendrogram produced by the algorithm. Moreover the hierarchical approach can cause difficulties in assessing the validity of the grouping [70]. From an opposite point of view, the use of different clustering techniques may lead to a variety of different groupings which can then be either analysed separately or merged

together in what is called 'consensus clustering'.

In Chapter 4, experiments based on the techniques mentioned above were reported. Several validity measures were examined as well, in an attempt to define the best number of clusters to consider rather than choosing *a priori* the proper number of groups. The results obtained were compared to previous work and discussed resorting to manual and visual characterisations. Six groups were then identified using a consensus clustering approach. These breast cancer classes appeared to have significant clinical meaning in terms of response to treatments and difference in survival rates. It has been observed that patients assigned to three of the six classes had the best survival, with up to 90% of women surviving at least 10 years. On the other side, the poorest survival was observed in women with HER2-positive breast cancer, where 30% died within four years. These six classes also appeared to be quite novel in literature, as two of them have not been emphasised yet.

This part of the project once again demonstrated the importance of using more than just a single clustering algorithm in breast cancer studies, as diverse methods usually produce different groupings. From a clinical perspective, it also served to move the field of breast tumours analysis from a single-technique approach toward a multi-technique one. However, the six identified groups could not be termed breast cancer 'phenotypes', as a consistent number of patients presented mixed class characteristics. It is a clear challenge for future work to investigate the proper characteristics of this big group of patients and their proper treatment.

- **A 'non-parametric' approach for supervised learning**.

  Supervised classification is a widely use approach in machine learning, and several different methods have been developed in recent years. One of the main advantages of supervised classification is that a set of rules is returned.

  To validate the breast cancer classes described above and in order to define a model for future patients classification, three different supervised classifiers were analysed

and applied to the same data. Results showed a surprising good performance of the naive Bayes classifier, even though one of its underlying assumptions (the normality distribution of variables) was strongly violated in the data. To overcome this limitation, a 'non-parametric' approach, similar to the naive Bayes and applicable independently from the covariates distribution was developed. This newly proposed algorithm is based on the ratio between areas under histograms representing variables distributions in each class, and the main decision criterion is related to how close to the median of a variable distribution in a specific class a particular datapoint is (the closer to the median, the higher the probability of being assigned to that particular class is). The method was also validated over three different data sets available from the Machine Learning Repository and results showed a better accuracy of the novel approach compared to the ones obtained by traditional techniques.

To complete the analysis, a comparison with Logistic Regression approach was performed, distinguishing between situations where the response variable is binary/boolean from those where a Multinomial model is needed (response variable may take more than two values). The technical details of the algorithms used have been presented in Chapter 5, together with all the results obtained.

- **CPU time for Affinity Propagation**.

  Several model-based algorithms have been proposed (see Section 2.5) in the last few years, but the most interesting one was the technique developed by Frey and Dueck called Affinity Propagation [60]. This method combines properties of both model-based and heuristic approaches to determine the optimal grouping. According to a more technical point of view, Affinity Propagation can be derived as the sum-product algorithm in a graphical model describing the mixture model [59]. The algorithm is claimed to be feasible with large data sets and faster than multiple runs of K-means [60].

  This algorithm was applied to several different cancer data sets, in particular to cuta-

neous melanoma and breast cancer case studies. Results confirmed what is already published in literature, but also gave novel insights, especially for the number of clusters to consider and for the clinical relevance of particular groups identified in the breast cancer data provided by the Nottingham City Hospitals. However, when considering the CPU time needed for the computation of affinity propagation, it was found that the relationship between the problem size and the CPU time was almost exponential, while a linear trend was observed for the same relationship for the K-means method. Based on this results, it was decided to exclude the affinity propagation from the clustering techniques considered in the proposed framework for the elucidation of core classes. Full details of the results are reported in Chapter 6.

- **Framework for classes identification**.

  One of the objectives of this thesis was to develop an algorithmic framework for the identification and characterisation of core stable groups in a collection of data (in order to determine their fundamental characteristics expressed by different groups). Cluster analysis may be thought as the discovery of distinct and non-overlapping sub-partitions within a larger population [130]. Using different algorithms and combining the results by a form of consensus clustering, core classes may be defined and characterised. Through clustering approaches, identifying labels may be assigned to objects and then used by supervised classification algorithms to learn classification rules and label new cases in a test set.

  This framework was presented in detail in Chapter 7, but the main structure is as follows. After pre-processing the data, dealing with all the missing information and computing the basic descriptive statistics, the application of several unsupervised clustering techniques was suggested. If the number of clusters is not known prior to commencing the analysis, many validation criteria are available to assess the goodness of the partitions obtained and to indicate the best number of groups to consider. Through biplots and boxplots clusters may be characterised and groups labels aligned to facilitate the analysis. Indices for the degree of agreement between

classification may be also computed to verify the results obtained so far and to have an indication about how likely a consensus between grouping will be. In the following step clusters are combined together and core classes are defined by considering only those cases which were assigned to the same group by different algorithms. Depending on the clustering results, this rule may be lighten by not considering all the clustering methods. The core classes obtained by this informal consensus may be again characterised by resorting to biplots and boxplots. To validate them, supervised classification techniques may be applied. Three different methods were suggested, which in general perform quite well with large data sets, with one of those being replaced by another approach developed in this work if particular hypotheses are not satisfied by the data under investigation.

This algorithmic framework was applied to a novel data set of histone markers provided by the Division of Molecular and Cellular Sciences in the School of Pharmacy at the University of Nottingham. Even though it is still a work in progress, the results obtained so far are very encouraging. The different classifications obtained by using different clustering methods have a high degree of agreement and the core classes identified by the consensus properly reflect the main characteristics of the data. They also appear to have distinct clinical qualities: they differ, indeed, in terms of survival rates and association with grading scores.

8.2 Potential clinical implications

In addition to the contributions already reported, this thesis and in particular the proposed framework may be also helpful for the development of a personalised breast cancer care. Nowadays, more and more research is focused on reaching this aim, as it is everyday more evident that each patient has his/her own requirements and needs specific treatment.

The discover of the six novel breast cancer classes (presented in Chapter 4) may be helpful for the future when a new patient may present to the hospital with the disease.

By analysing the results of the first biological tests and relate them with the principal characteristics of the six groups, it might be possible to categorise this new patient in a specific class. Then, the most powerful treatment could be directly given, thus reducing the pain for the patient and the clinical costs.

Moreover, as the proposed framework is aimed to be applicable to any kind of problem, it might be useful to train a small number of clinicians so that they can directly use the framework and find different possible categories of the disease they are investigating.

8.3 Future work

Several directions of research may be followed for further improvements of this thesis work. Some of them are reported below, including those which are currently being carried out.

Clustering algorithms

For cluster analysis, several unsupervised methods were used in this study to enhance groups in breast cancer data. Different methods have their own advantages and disadvantages for specific clustering criteria. Therefore, it may be useful to investigate the combination of other clustering algorithms, which may have different optimisation criteria. Model-based or density-based clustering algorithms could be examined, instead of hierarchical and partitional methods. As already highlighted in this thesis, different algorithms produce different clusters and a 'perfect' technique suitable for any kind of data has not been developed yet.

Clustering initialisation techniques

For those clustering techniques that do not use actual data points as cluster centres, the initial set of centroids is of particular importance. Usually a random assignment is performed at the beginning, thus resulting in minor differences in clustering groupings in different runs of the same algorithm. In order to reproduce the results, the approach

followed in this work was to set the initial cluster centres using the first run of a hierarchical algorithm. However, various techniques have been proposed for clustering initialisation [4] and could be used as a future research. In particular, there is still place for the investigation of a method which can set 'the appropriate' initial cluster centres independently from the type of data being studied. This could make the clustering algorithms working more efficiently and perhaps leading to different and more convincing results for the fuzzy c-means method applied to the same breast cancer data set.

Distances for clustering approaches

In all clustering techniques used throughout this work the Euclidean distance was used to measure the separation of data in multi-dimensional space and to represent dissimilarities. However, if the shape of clusters is far from being spherical, it may result in an inappropriate clustering. For this reason, other distance measures and other forms of dissimilarities may be employed. One minus Pearson correlation, Canberra or Manhattan distances may be used instead of the Euclidean one.

Validity indices

To assess clustering results and to define the best number of groups to consider in the analysis, several validity indices were used. As reported in [126, 183] many different validity indices have been proposed in recent years. For the clustering analysis described in this work, only six indices, that were already implemented in R, have been used. However, one direction for future research could be the investigation of different indices or even the use of clustering algorithms which have an internal criterion to verify the stability of the returned clusters.

Investigation of 'not-classified' patients

By a form of consensus clustering six core breast cancer classes have been defined in this work. However, as highlighted in Chapter 4, not all the available patients were classified in one of these six groups. A very important future research will be to define a proper classification for those patients in order to help doctors give them more accurate

prognoses, as well as targeting patients with more specialised treatments. This represents a big challenge for future work, as finding the proper cure for each patient will decrease hospital costs as well as the patient pain.

Getting new patients

One of the strategies that may be followed to achieve the previous goal might be to increase the number of available patients. This could be done by retrieving medical records or by performing again the same biological analyses in order to recover some missing data. It would be interesting to investigate if it could be feasible to combine different sources of data by merging studies, from different research groups, in which data have been collected using very similar protocols.

Different supervised learning

Different supervised classification techniques have been used to verify known groupings. Following valuable comments and suggestions raised during conferences presentations, a direction for future research could be the investigation of support vector machines as supervised learning. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximises the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are 'pushed up against' the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring datapoints of both classes, since in general the larger the margin the lower the generalization error of the classifier. SVMs were firstly developed by Cortes and Vapnik (1995) for binary classification [32].

Characterisation of Affinity Propagation groups

As shown in Chapter 6, the Affinity Propagation (AP) algorithm combines properties of both model-based and heuristic approaches. This method was successfully applied to

several cancer studies and interesting results, compared with those of original works, were obtained. As a further investigation, it would be interesting to characterise the AP groups on the basis of gene expressions and to relate the clinical outcomes of these groups with the ones obtained by standard techniques in the original papers.

R package for Affinity Propagation algorithm

The Affinity Propagation was recently developed by Frey and Dueck [60]. In the Supplementary Material on Science website, authors provided the MatLab code (Version 7, Release 14) for the implementation of this deterministic algorithm. As members of a School of Computer Science there is a particular interest in developing a free R package for the implementation of Affinity Propagation, as from many parts in the scientific community this request was raised. At time of writing (November 2009), a first version working only with small and trivial data sets was developed.

Complete the analysis on histone markers

In this thesis a guideline to detect core classes in a data set was proposed and validated over a novel case study on histone markers for breast cancer. As already mentioned, this is still an ongoing work, and the next steps will be concentrated on both technical and clinical aspects. For the former, a tri-dimensional visualisation of classes distribution will be provided using the first three principal components, in order to verify the compactness of groups and their separation in 3D space. Concerning the clinical aspect, more information will be provided by researcher in the School of Pharmacy and the association between the identified core classes and these further clinical variables will be studied. As the cohort of patients under investigation is the same for both studies, it would be very interesting to investigate a possible relation between the six breast cancer classes emphasised by the consensus methodology described in Chapter 4 and the ones identified in this part of the work.

8.4 Dissemination

The research work reported in this thesis has been used in various conference and journal papers as well as several internal and international talks. What follows is a list of publications and presentations derived from this work, together with a reference to the chapter in which the topic is covered.

8.4.1 Journal papers

In submission

D. Soria, J.M. Garibaldi, F. Ambrogi, E. Biganzoli and I.O. Ellis, *A 'Non-Parametric' Version of the Naive Bayes Classifier*, submitted to Data & Knowledge Engineering, 2009. (Chapter 5)

E. Biganzoli, D. Coradini, F. Ambrogi, J.M. Garibaldi, P.J.G. Lisboa, D. Soria, A.R. Green, M. Pedriali, M. Piantelli, P. Querzoli, R. Demicheli, P. Boracchi, I. Nenci, I.O. Ellis, S. Alberti, *p53 Status Identifies Two Subgroups of Triple-Negative Breast Cancers with Distinct Prognosis*, submitted to Breast Cancer Research, 2009. (Chapter 4)

Accepted

D. Soria, J.M. Garibaldi, F. Ambrogi, A.R. Green, D.G. Powe, E.A. Rakha, R.D. Macmillan, R.W. Blamey, G.R. Ball, P.J.G. Lisboa, T.A. Etchells, P. Boracchi, E. Biganzoli and I.O. Ellis, *A Methodology to Identify Consensus Classes from Clustering Algorithms Applied to Immunohistochemical Data from Breast Cancer Patients*, accepted for publication in Computers in Biology and Medicine, 2009. (Chapter 4)

D. Soria, F. Ambrogi, E. Raimondi, P. Boracchi, J.M. Garibaldi and E. Biganzoli, *Cancer Profiles by Affinity Propagation*, Special issue of the International Journal of Knowledge Engineering and Soft Data Paradigms (IJKESDP), 1(3):195-215, 2009. (Chapter 6)

S.E. Elsheikh, A.R. Green, E.A. Rakha, D.G. Powe, R.A. Ahmed, H.M. Collins, D. Soria, J.M. Garibaldi, C.E. Paish, A.A. Ammar, M.J. Grainge, G.R. Ball, M.K. Abdelghany, L. Martinez-Pomares, D.M. Heery and I.O. Ellis, *Global Histone Marks in Breast Cancer Correlate with Tumour Phenotypes, Prognostic Factors and Patient Outcome*, Cancer Research, 69:3802-3809, 2009.

8.4.2 Conference papers

D.Soria, F. Ambrogi, P. Boracchi, J.M. Garibaldi, E. Biganzoli, *Application of Affinity Propagation to a Large Breast Cancer Data Set*, Proceedings of the Conference on Statistical methods for the analysis of large data sets (SIS 2009), 531-534, Pescara, Italy, 23 - 25 September 2009. (Chapter 6)

F. Ambrogi, E. Raimondi, D. Soria, P. Boracchi, E. Biganzoli, *Cancer Profiles by Affinity Propagation*, in the Proceedings of the Seventh International Conference on Machine Learning and Applications (ICMLA08), IEEE Computer Society, 650-655, San Diego, US, 11 - 13 December 2008. (Chapter 6)

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *A Comparison of Three Different Methods for Classification of Breast Cancer Data*, in the Proceedings of the Seventh International Conference on Machine Learning and Applications (ICMLA08), IEEE Computer Society, 619-624, San Diego, US, 11 - 13 December 2008. (Chapter 5)

D. Soria, J.M. Garibaldi, F. Ambrogi, P.J.G. Lisboa, P. Boracchi, E. Biganzoli, *Clustering Breast Cancer Data by Consensus of Different Validity Indices*, in the Proceedings of the 4th International Conference on Advances in Medical, Signal and Information Processing (MEDSIP), Santa Margherita Ligure, Italy, 14 - 16 July 2008. (Chapter 4)

8.4.3 Presentations

D.Soria, F. Ambrogi, P. Boracchi, J.M. Garibaldi, E. Biganzoli, *Application of Affinity Propagation to a Large Breast Cancer Data Set*, (oral presentation) at the Conference on Statistical methods for the analysis of large data sets (SIS 2009), Pescara, Italy,

24th September 2009.

D. Soria, F. Ambrogi, J.M. Garibaldi, P. Boracchi and E. Biganzoli, *Application of Affinity Propagation on Breast Cancer Data Sets*, (oral presentation) at the 23rd European Conference on Operational Research (EURO XXIII), Bonn, Germany, 6th July 2009.

D. Soria, F. Ambrogi, J.M. Garibaldi, P. Boracchi and E. Biganzoli, *Application of Affinity Propagation on Breast Cancer Data Sets*, (oral presentation) at Intelligent Modelling and Analysis Research Group seminar, School of Computer Science, University of Nottingham, Nottingham, UK, 23rd June 2009.

D. Soria, F. Ambrogi, J.M. Garibaldi, P. Boracchi and E. Biganzoli, *Application of Affinity Propagation on Breast Cancer Data Sets*, (oral presentation) at the 5th BIOPTRAIN Workshop, Firenze, Italy, 10th June 2009.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Bioinformatics Analysis of Breast Cancer Data*, (oral presentation) at the 4th BIOPTRAIN Workshop, Innsbruck, Austria, 12th January 2009.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *A Comparison of Three Different Methods for Classification of Breast Cancer Data*, (oral presentation) at the Seventh International Conference on Machine Learning and Applications (ICMLA08), San Diego, US, 13th December 2008.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Classification Techniques for Breast Cancer Data*, (oral presentation) at the Institute of Computing Science School Seminar, Poznan University of Technology, Poznan, Poland, 25th November 2008.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Classification Techniques for Breast Cancer Data*, (oral presentation) at the Mini EURO Conference on Computational Biology, Bioinformatics and Medicine (Mini EURO-CCBBM), Rome, Italy, 17th September 2008.

D. Soria, J.M. Garibaldi, F. Ambrogi, P.J.G. Lisboa, P. Boracchi and E. Biganzoli, *Clustering Breast Cancer Data by Consensus of Different Validity Indices*, (oral presentation) at the Fourth International Conference on Advances in Medical, Signal and

Information Processing (MEDSIP 2008), Santa Margherita Ligure, Italy, 14th July 2008.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Identification of Key Breast Cancer Phenotypes*, (oral presentation) at the Automated Scheduling Optimisation and Planning Research Group seminar, School of Computer Science, University of Nottingham, Nottingham, UK, 1st November 2007.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Comparison of Clustering Techniques for Breast Cancer Data*, (oral presentation) at the 22nd European Conference on Operational Research (EURO XXII), Prague, Czech Republic, 11th July 2007.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Identification of Sub-Populations in Breast Cancer Data through Different Clustering Algorithms*, (oral presentation) at the third BIOPTRAIN internal meeting, School of Computer Science, University of Nottingham, Nottingham, UK, 9th January 2007.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Identification of Sub-Populations in Breast Cancer Data through Different Clustering Algorithms*, (oral presentation) at the Automated Scheduling Optimisation and Planning Research Group seminar, School of Computer Science, University of Nottingham, Nottingham, UK, 13th December 2006.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Progress Report*, (oral presentation) at the second BIOPTRAIN internal meeting, School of Computer Science, University of Nottingham, Nottingham, UK, 12th July 2006.

D. Soria, J.M. Garibaldi, E. Biganzoli and I.O. Ellis, *Bioinformatics Analysis of Breast Cancer Data*, (oral presentation) at the first BIOPTRAIN internal meeting, School of Computer Science, University of Nottingham, Nottingham, UK, 2nd March 2006.

References

[1] D.M. Abd El-Rehim, G. Ball, S.E. Pinder, E. Rakha, C. Paish, J.F. Robertson, D. Macmillan, R.W. Blamey, and I.O. Ellis. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. Journal of Cancer*, 116:340–350, 2005.

[2] D.M. Abd El-Rehim, S.E. Pinder, C.E. Paish, J.A. Bell, R.W. Blamey, J.F. Robertson, R.I. Nicholson, and I.O. Ellis. Expression of luminal and basal cytokeratins in human breast carcinoma. *Journal of Pathology*, 203:661–671, 2004.

[3] D.M. Abd El-Rehim, S.E. Pinder, C.E. Paish, J.A. Bell, R.S. Rampaul, R.W. Blamey, J.F. Robertson, R.I. Nicholson, and I.O. Ellis. Expression and co-expression of the members of the epidermal growth factor receptor (EGFR) family in invasive breast carcinoma. *Br. J. Cancer*, 91(8):1532–1542, 2004.

[4] M. Al-Daoud and S. Roberts. New methods for the initialisation of clusters. *Pattern Recognition Letters*, 17(5):451–455, 1996.

[5] F. Ambrogi, E. Biganzoli, P. Querzoli, S. Ferretti, P. Boracchi, S. Alberti, E. Marubini, and I. Nenci. Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods. *Clinical Cancer Research*, 12(3):781–790, 2006.

[6] Cluster Analysis. Copyright StatSoft, Inc., 1984-2004. `http://www.statsoft.com/textbook/stcluan.html`.

[7] M.R. Anderberg. *Cluster Analysis for Applications*. Monographs and Textbooks on Probability and Mathematical Statistics. Academic Press, Inc., New York, 1973.

[8] A. Asuncion and D.J. Newman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007. `http://archive.ics.uci.edu/ml/`.

[9] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[10] A. Bellaachia and E. Guven. Predicting breast cancer survivability using data mining techniques. *Scientific Data Mining Workshop, in Conjunction with the 2006 SIAM Conference on Data Mining*, 2006.

[11] J.C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–73, 1974.

[12] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York edition, 1981.

[13] J.C. Bezdek, R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10:191–203, 1984.

[14] C.M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer edition, 2006.

[15] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

[16] S. Borman. The Expectation Maximisation algorithm: A short tutorial. Online at: `http://www.seanborman.com/publications/EM_algorithm.pdf`, 2004.

[17] R.R. Bouckaert. Naive bayes classifiers that perform well with continuous variables. In *Proceedings of the 17th Australian Conference on AI (AI04)*. Berlin: Springer, 2004.

[18] J.D. Brenton, L.A. Carey, A.A. Ahmed, and C. Caldas. Molecular classification and molecular forecasting of breast cancer: Ready for clinical application? *J Clin Oncol*, 23:7350–7360, 2005.

[19] E.K. Burke, E. Hart, G. Kendall, J. Newall, P. Ross, and S. Schulenburg. Hyper-heuristics: An emerging direction in modern search technology. *Handbook of Metaheuristics (F. Glover and G. Kochenberger, eds.), Kluwer*, pages 457–474, 2003.

[20] R.B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communs statist*, 3:1–27, 1974.

[21] G. Callagy, E. Cattaneo, Y. Daigo, L. Happerfield, L. Bobrow, P. Pharoah, and C. Caldas. Molecular classification of breast carcinomas using tissue microarrays. *Diagn Mol Pathol*, 12:27–34, 2003.

[22] S. Calza, P. Hall, G. Auer, J. Björhle, S. Klaar, U. Kronenwett, E.T. Liu, L. Miller, A. Ploner, J. Smeds, J. Bergh, and Y. Pawitan. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Research*, 8:R34, 2006.

[23] X. Cao, K.B. Maloney, and V. Brusic. Data mining of cancer vaccine trials: A bird's-eye view. *Immunome Research*, 4:7, 2008.

[24] R.D. Cardiff, U. Wagner, and L. Henninghausen. Mammary cancer in humans and mice: A tutorial for comparative pathology. *Vet Pathol*, 38(4):357–358, 2001.

[25] L.A. Carey, E.C. Dees, L. Sawyer, L. Gatti, D.T. Moore, F. Collichio, D.W. Ollila, C.I. Sartor, M.L. Graham, and C.M. Perou. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clin Cancer Res*, 13:2329–2334, 2007.

[26] G.A. Carpenter and S. Grossberg. ART2: Stable self-organization of pattern recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930, 1987.

[27] National Center for Biotechnology Information. Microarrays: Chipping away at the mysteries of science and medicine, 2007. `http://www.ncbi.nlm.nih.gov/About/primer/microarrays`.

[28] G. Cimoli, D. Malacarne, R. Ponassi, M. Valenti, S Alberti, and S. Parodi. Meta analysis of the role of p53 status in isogenic systems tested for sensitivity to cytotoxic anti-neoplastic drugs. *Biochem Biophys Acta*, 1705:103–120, 2004.

[29] S. Cleator, W. Heller, and R.C. Coombes. Triple-negative breast cancer: Therapeutic options. *Lancet Oncol*, 8:235–244, 2007.

[30] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[31] J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 1968.

[32] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[33] P. de Cremoux, A. Vincent Salomon, S. Liva, R. Dendale, B. Bouchind'homme, E. Martin, X. Sastre-Garau, H. Magdelenat, A. Fourquet, and T. Soussi. p53 mutation as a genetic trait of typical medullary breast carcinoma. *J. Natl. Cancer Inst.*, 91(7):641–643, 1999.

[34] D. Delen, G. Walker, and A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2):113–127, 2005.

[35] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[36] R. Dent, M. Trudeau, K.I. Pritchard, W.M. Hanna, H.K. Kahn, C.A. Sawka, L.A. Lickley, E. Rawlinson, P. Sun, and S.A. Narod. Triple-negative breast cancer: Clinical features and patterns of recurrence. *Clin Cancer Res*, 13:4429–4434, 2007.

[37] S. Detre, G. Saclani Jotti, and M. Dowsett. A "quickscore" method for immunohistochemical semiquantitation: Validation for oestrogen receptor in breast carcinomas. *J Clin Pathol*, 48:876–878, 1995.

[38] A. Di Leo, M. Tanner, C. Desmedt, M. Paesmans, F Cardoso, V. Durbecq, S. Chan, T. Perren, M. Aapro, C. Sotiriou, M.J. Piccart, D. Larsimont, and J. Isola. p-53 gene mutations as a predictive marker in a population of advanced breast cancer patients randomly treated with doxorubicin or docetaxel in the context of a phase III clinical trial. *Ann Oncol*, 18:997–1003, 2007.

[39] R. Diallo-Danebrock, E. Ting, O. Gluz, A. Herr, S. Mohrmann, H. Geddert, A. Rody, K.L. Schaefer, S.E. Baldus, A. Hartmann, P.J. Wild, M. Burson, H.E. Gabbert, U. Nitz, and C. Poremba. Protein expression profiling in high-risk breast cancer patients treated with high-dose or conventional dose-dense chemotherapy. *Clin Cancer Res*, 13:488–497, 2007.

[40] M. Dolled-Filhart, L. Rydén, M. Cregger, K. Jirström, M. Harigopal, R.L. Camp, and D.L. Rimm. Classification of breast cancer using genetic algorithms and tissue microarrays. *Clin Cancer Res*, 12:6459–6468, 2006.

[41] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *ICML*, pages 194–202. Morgan Kaufmann, 1995.

[42] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):research0036.1–0036.21, 2002.

[43] D. Dueck, B.J. Frey, N. Jojic, V. Jojic, G. Giaever, A. Emili, G. Musso, and R. Hegele. *Constructing Treatment Portfolios Using Affinity Propagation*, volume 4955 of *Lecture Notes in Computer Science*, pages 360–371. Springer Berlin / Heidelberg, 2008.

[44] J.C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1974.

[45] J.C. Dunn. Well separated clusters and optimal fuzzy partition. *Journal of Cybernetics*, 4:95–104, 1974.

[46] A.W.F. Edwards and L.L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21(2):362–375, 1965.

[47] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–14868, 1998.

[48] I.O. Ellis, M. Galea, N. Broughton, A. Locker, R.W. Blamey, and C.W. Elston. Pathological prognostic factors in breast cancer. II. histological type. relationship with survival in a large study with long-term follow-up. *Histopathology*, 20:479–489, 1992.

[49] I.O. Ellis, S.E. Pinder, A.H. Lee, and C.W. Elston. A critical appraisal of existing classification systems of epithelial hyperplasia and in situ neoplasia of the breast with proposals for future methods of categorization: Where are we going? *Semin Diagn Pathol*, 16:202–208, 1999.

[50] S.E. Elsheikh, A.R. Green, M.B.K. Lambros, N.C. Turner, M.J. Grainge, D. Powe, I.O. Ellis, and J.S. Reis-Filho. FGFR1 amplification in breast carcinomas: A chromogenic in situ hybridisation analysis. *Breast Cancer Research*, 9:R23, 2007.

[51] S.E. Elsheikh, A.R. Green, E.A. Rakha, D.G. Powe, R.A. Ahmed, H.M. Collins, D. Soria, J.M. Garibaldi, C.E. Paish, A.A. Ammar, M.J. Grainge, G.R. Ball, M.K. Abdelghany, L. Martinez-Pomares, D.M. Heery, and I.O. Ellis. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. *Cancer Research*, 69:3802–3809, 2009.

[52] T.A. Etchells and P.J.G. Lisboa. Rule extraction from neural networks: a practical and efficient approach. *IEEE Transactions on Neural Networks*, 17(2):374–384, 2006.

[53] B.S. Everitt. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2002.

[54] I.W. Evett and E.J. Spiehler. Rule induction in forensic science. In *KBS in Goverment*, pages 107–118. Online Publications, 1987.

[55] I.S. Fentiman. The dilemma of in situ carcinoma of the breast. *Int J Clin Pract*, 55(10):680–683, 2001.

[56] I.S. Fentiman. Timing of surgery for breast cancer. *Int J Clin Pract*, 56(3):188–190, 2002.

[57] V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 418– 426, 2003.

[58] C. Fraley and A.E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[59] B.J. Frey and D. Dueck. Mixture modeling by affinity propagation. *Advances in Neural Information Processing Systems*, 2006.

[60] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[61] H.P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178, 1967.

[62] L.G. Fulford, J.S. Reis-Filho, K. Ryder, C. Jones, C.E. Gillett, A. Hanby, D. Easton, and S.R. Lakhani. Basal-like grade III invasive ductal carcinoma of the breast: Patterns of metastasis and long-term survival. *Breast Cancer Res*, 9:R4, 2007.

[63] K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[64] M.H. Galea, R.W. Blamey, C.E. Elston, and I.O. Ellis. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat*, 22:207–219, 1992.

[65] M. Garcia-Closas and S. Chanock. Genetic susceptibility loci for breast cancer by estrogen receptor status. *Clin Cancer Res*, 14:8000–8009, 2008.

[66] E. Garrett-Mayer and G. Parmigiani. Clustering and classification methods for gene expression data analysis. *Johns Hopkins University, Dept.of Biostatistics Working Papers, Johns Hopkins University*, 2004. `http://www.bepress.com/jhubiostat/`.

[67] I. Gath and A.B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–781, 1989.

[68] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A*, 97(22):12079–12084, 2000.

[69] I. Gitman and M.D. Levine. An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique. *IEEE Transactions on Computers*, C-19(7):583–593, 1970.

[70] D.R. Goldstein, G. Debashis, and E.M. Conlon. Statistical issues in the clustering of gene expression data. *Statistica Sinica*, 12:219–240, 2002.

[71] A.D. Gordon. *Classification: Methods for the Exploratory Analysis of Multivariate Data*. New York: Chapman and Hall, 1981.

[72] B. Gusterson. Do 'basal-like' breast cancers really exist? *Nature Reviews Cancer*, 2008.

[73] S.J. Haberman. Generalized residuals for log-linear models. In *Proceedings of the 9th International Biometrics Conference*, pages 104–122, 1976.

[74] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[75] F.E. Harrell Jr., K.L. Lee, and D.B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[76] L.N. Harris, G. Broadwater, and N.U. Liu. Molecular subtypes of breast cancer in relation to paclitaxel response and outcomes in women with metastatic disease: Results from CALGB 9342. *Breast Cancer Res*, 8:R66, 2006.

[77] J.A. Hartigan. *Clustering Algorithms*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1975.

[78] J.A. Hartigan and M.A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[79] J.L. Haybittle, R.W. Blamey, C.W. Elston, J. Johnson, P.J. Doyle, F.C. Campbell, R.I. Nicholson, and K. Griffiths. A prognostic index in primary breast cancer. *Br J Cancer*, 45:361–366, 1982.

[80] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2 edition, 1998.

[81] World Health Organization (February 2009). Fact sheet no. 297: Cancer. `http://www.who.int/mediacentre/factsheets/fs297/en/index.html`, 2009.

[82] World Health Organization International Agency for Research on Cancer. World cancer report. `http://www.iarc.fr/en/publications/pdfs-online/wcr/index.php`, 2008.

[83] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[84] Z. Hu, C. Fan, D.S. Oh, J.S. Marron, X. He, B.F. Qaqish, C. Livasy, L.A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M.G. Ewend, L.R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Ruiz Orrico, D. Dreher, J.P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J.F. Quackenbush, M.J. Ellis, O.I. Olopade, P.S. Bernard, and C.M. Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7:96, 2006.

[85] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[86] J.E. Jackson. *A User's Guide to Principal Components*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1991.

[87] J. Jacquemier, C. Ginestier, J. Rougemont, V.-J. Bardou, E. Charafe-Jauffret, J. Geneix, J. Adélaïde, A. Koki, G. Houvenaeghel, J. Hassoun, D. Maraninchi, P. Viens, D. Birnbaum, and F. Bertucci. Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. *Cancer Res*, 65:767–779, 2005.

[88] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall advanced reference series, Prentice-Hall. Englewood Cliffs, NJ, USA, 1988.

[89] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[90] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and data Engineering*, 16(11):1370–1386, 2004.

[91] G.H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[92] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York, 2002.

[93] J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Hoboken, N.J. : Wiley-Interscience, 2nd edition, 2002.

[94] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[95] A.V. Kapp, S.S. Jeffrey, A. Langerød, A.-L. Børresen-Dale, W. Han, D.-Y. Noh, I.R.K. Bukholm, M. Nicolau, P.O. Brown, and R. Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(231), 2006.

[96] T. Käster, V. Wendt, and G. Sagerer. *Comparing Clustering Methods for Database Categorization in Image Retrieval*, pages 228–235. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2003.

[97] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1990.

[98] P. Kellam, X. Liu, N. Martin, C. Orengo, S. Swift, and A. Tucker. Comparing, contrasting and combining clusters in viral gene expression data. In *Proceedings of 6th Workshop on Intelligent Data Analysis in Medicine*, 2001.

[99] J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M.J. Mihatsch, G. Sauter, and O.-P. Kallionimeni. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4:844–847, 1998.

[100] E. Korsching, J. Packeisen, K. Agelopoulos, M. Eisenacher, R. Voss, J. Isola, P.J. van Diest, B. Brandt, W. Boecker, and H. Buerger. Cytogenetic alterations and cytokeratin expression patterns in breast cancer: Integrating a new model of breast

differentiation into cytogenetic pathways of breast carcinogenesis. *Lab Invest*, 82:1525–1533, 2002.

[101] S.B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[102] W.J. Krzanowski and Y.T. Lai. A criterion for determining the number of clusters in a data set. *Biometrics*, 44:23–34, 1985.

[103] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[104] D.A. Kulesh, D.R. Clive, D.S. Zarlenga, and J.J. Greene. Identification of interferon-modulated proliferation-related cdna sequences. *Proc Natl Acad Sci USA*, 84:8453–8457, 1987.

[105] M. Laakso, M. Tanner, J. Nilsson, T. Wiklund, B. Erikstein, P. Kellokumpu-Lehtinen, P. Malmström, N. Wilking, J. Bergh, and J. Isola. Basoluminal carcinoma: A new biologically and prognostically distinct entity between basal and luminal breast cancer. *Clinical Cancer Research*, 12:4185–4191, 2006.

[106] S.R. Lakhani, M.J. van de Vijver, J. Jacquemier, T.J. Anderson, P.P. Osin, L. McGuffog, and D.F. Easton. The pathology of familial breast cancer: Predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, HER-2, and p53 in patients with mutations in BRCA1 and BRCA2. *J Clin Oncol*, 20:2310–2318, 2002.

[107] N. Lama, F. Ambrogi, L. Antolini, P. Boracchi, and E. Biganzoli. Some issues and perspective in microarray data analysis in breast cancer: the need for an integrated research. In *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance*, 2004.

[108] A. Langerød, H. Zhao, Ø. Borgan, J.M. Nesland, I.R.K. Bukholm, T. Ikdahl, R. Kåresen, A.-L. Børresen-Dale, and S.S. Jeffrey. TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res*, 9:R30, 2007.

[109] M. Leone, Sumedha, and M. Weigt. Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics*, 23:2708–2715, 2007.

[110] A.V. Lukashin and R. Fuchs. Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17:405–414, 2001.

[111] R.D. Macmillan. Screening women with a family history of breast cancer – results from the British Familial Breast Cancer Group. *European Journal of Surgical Oncology*, 26:149–152, 2000.

[112] J.B. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley*, pages 281–297, 1967.

[113] D. Maglott, J. Ostell, K.D. Pruitt, and T. Tatusova. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, Database Issue:D54–D58, 2005.

[114] J.H. Maindonald and W.J. Braun. *Data Analysis and Graphics Using R - An Example-Based Approach*. Cambridge University Press, 2003.

[115] N.A. Makretsov, D.G. Huntsman, T.O. Nielsen, E. Yorida, M. Peacock, M.C.U. Cheang, S.E. Dunn, M. Hayes, M. van de Rijn, C. Bajdik, and C. Blake Gilks. Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. *Clin Cancer Res*, 10:6143–6151, 2004.

[116] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[117] Merck Manual of Diagnosis and Therapy. Breast Disorders: Cancer, 2008. `http://www.merck.com/mmhe/sec22/ch251/ch251f.html`.

[118] F.H.C. Marriot. Practical problems in a method of cluster analysis. *Biometrics*, 27(3):501–514, 1971.

[119] B. Matharoo-Ball, L. Ratcliffe, L. Lancashire, S. Ugurel, A.K. Miles, D.J. Weston, R. Rees, D. Schadendorf, G. Ball, and C.S. Creaser. Diagnostic biomarkers differentiating metastatic melanoma patients from healthy controls identified by an integrated MALDI-TOF mass spectrometry/bioinformatic approach. *Proteomics Clin. Appl.*, 1(6):605–620, 2007.

[120] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

[121] K.S. McCarty Jr., L.S. Miller, E.B. Cox, J. Konrath, and K.S. McCarty Sr. Estrogen receptor analyses. Correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Arch Pathol Lab Med*, 109:716–721, 1985.

[122] R.A. McClelland, P. Finlay, K.J. Walker, D. Nicholson, J.F.R. Robertson, R.W. Blamey, and R.I. Nicholson. Automated quantitation of immunocytochemically localized estrogen receptors in human breast cancer. *Cancer Res*, 50:3545–3550, 1990.

[123] M. Meilă and D. Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42:9–29, 2001.

[124] S. Ménard, P. Casalini, G. Tomasic, S. Pilotti, N. Cascinelli, R. Bufalino, F. Perrone, C. Longhi, F. Rilke, and M.I. Colnaghi. Pathobiologic identification of two distinct breast carcinoma subsets with diverging clinical behaviors. *Breast Cancer Res Treat*, 55(2):169–77, 1999.

[125] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458):488–492, 2005.

[126] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[127] G.W. Milligan and M.C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21:441–458, 1986.

[128] T.M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[129] T.M. Mitchell. Generative and discriminative classifiers: Naive bayes and logistic regression. `http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf`, 2005.

[130] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118, 2003.

[131] A.Y. Ng and M.I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems (NIPS)*, 14, 2002.

[132] T.O. Nielsen, F.D. Hsu, K. Jensen, M Cheang, G. Karaca, Z. Hu, T. Hernandez-Boussard, C. Livasy, D. Cowan, L. Dressler, L.A. Akslen, J. Ragaz, A.M. Gown, C. Blake Gilks, M. van de Rijn, and C.M. Perou. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res*, 10:5367–5374, 2004.

[133] Department of Measurement and Health Information Systems:. World Health Statistics 2007. World Health Organization, Geneva, Switzerland, 2007.

[134] E. Ozcan, B. Bilgin, and E.E. Korkmaz. Hill climbers and mutational heuristics in hyperheuristics. In *Lecture Notes in Computer Science, Springer-Verlag, The 9th International Conference on Parallel Problem Solving From Nature*, pages 202–211, 2006.

[135] E. Ozcan, B. Bilgin, and E.E. Korkmaz. A comprehensive analysis of hyper-heuristics. *Intelligent Data Analysis*, 12(1):3–23, 2008.

[136] H. Ozcelik, D. Pinnaduwage, S.B. Bull, and I.L. Andrulis. Type of TP53 mutation and ERBB2 amplification affects survival in node-negative breast cancer. *Breast Cancer Res Treat*, 105:255–265, 2007.

[137] C.M. Perou, T. Sørlie, M.B. Eisen, M. Van De Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lonning, A.L. Børresen-Dale, P.O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[138] J.R. Pollack, T. Sørlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Børresen-Dale, and P.O. Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99:12963–12968, 2002.

[139] L. Pusztai, M. Ayers, J. Stec, E. Clark, K. Hess, D. Stivers, A. Damokosh, N. Sneige, T.A. Buchholz, F.J. Esteva, B. Arun, M. Cristofanilli, D. Booser, M. Rosales, V. Valero, C. Adams, G.N. Hortobagyi, and W.F. Symmans. Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. *Clin Cancer Res*, 9:2406–2415, 2003.

[140] P. Querzoli, S. Ferretti, G. Albonico, E. Magri, D. Scapoli, M. Indelli, and I. Nenci. Application of quantitative analysis to biologic profile evaluation in breast cancer. *Cancer*, 76(12):2510–2517, 1995.

[141] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Los Altos, California, 1993.

[142] R. Radhakrishnan, M. Solomon, K. Satyamoorthy, L.E. Martin, and M.W. Lingen. Tissue microarray – a high-throughput molecular analysis in head and neck cancer. *J Oral Pathol Med*, 37:166–176, 2008.

[143] E.A. Rakha, M.E. El-Sayed, A.R. Green, A.H.S. Lee, J.F. Robertson, and I.O. Ellis. Prognostic markers in triple-negative breast cancer. *Cancer*, 109:25–32, 2007.

[144] E.A. Rakha, M.E. El-Sayed, A.H.S. Lee, C.W. Elston, M.J. Grainge, Z. Hodi, R.W. Blamey, and I.O. Ellis. Prognostic significance of nottingham histologic grade in invasive breast carcinoma. *J Clin Oncol*, 26(19):3153–3158, 2008.

[145] E.A. Rakha, T.C. Putti, D.M. Abd El-Rehim, C. Paish, A.R. Green, A.H. Lee, J.F. Robertson, and I.O. Ellis. Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *Journal of Pathology*, 208:495–506, 2006.

[146] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[147] M. Reddy and R. Given-wilson. Screening for breast cancer. *Surgery*, 22(7):155–160, 2004.

[148] M.R. Rezaee, B.B.F. Lelieveldt, and J.H.C. Reiber. A new cluster validity index for the fuzzy *C*-mean. *Pattern Recognition Letters*, 19(3-4):237–246, 1998.

[149] P. Ross. Hyper-heuristics, Search Methodologies: Introductory Tutorials. Optimization and Decision Support Techniques (E. K. Burke and G. Kendall, eds.), Springer, 2005.

[150] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, 1987.

[151] R. Rouzier, C.M. Perou, W.F. Symmans, N. Ibrahim, M. Cristofanilli, K. Anderson, K.R. Hess, J. Stec, M. Ayers, P. Wagner, P. Morandi, C. Fan, I. Rabiul, J.S. Ross, G.N. Hortobagyi, and L. Pusztai. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res*, 11:5678–5685, 2005.

[152] P. Royston. Algorithm AS 181: The W test for normality. *Applied Statistics*, 31:176–180, 1982.

[153] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[154] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397, 1971.

[155] D. Shalon, S.J. Smith, and P.O. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, 6:639–645, 1996.

[156] J.P. Siebert. Vehicle recognition using rule based methods. *Turing Institute Research Memorandum TIRM-87-018*, 1987.

[157] R.M. Simon, E.L. Korn, L.M. McShane, M.D. Radmacher, G.W. Wright, and Y. Zhao. *Design and Analysis of DNA Microarray Investigations*. Springer, 2004.

[158] T. Sørlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. Van De Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P. Eystein Lonning, and A.L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98:10869–10874, 2001.

[159] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J.S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C.M. Perou, P.E. Lonning, P.O. Brown, A.L. Børresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100:8418–8423, 2003.

[160] T. Sørlie, Y. Wang, C. Xiao, H. Johnsen, B. Naume, R.R. Samaha, and A.L. Børresen-Dale. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: Gene expression analyses across three different platforms. *BMC Genomics*, 7:127, 2006.

[161] C. Sotiriou, S.-Y. Neo, L.M. McShane, E.L. Korn, P.M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A.L. Harris, and E.T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100:10393–10398, 2003.

[162] H. Sun, S. Wang, and Q. Jiang. FCM-based model selection algorithms for determining the number of clusters. *Pattern Recognition*, 37(10):2027–2037, 2004.

[163] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam. Consensus clustering and functional interpretation of gene-expression data. *Genome Biology*, 5:R94, 2004.

[164] T. Tang, N. François, A. Glatigny, N. Agier, M.H. Mucchielli, L. Aggerbeck, and H. Delacroix. Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment. *Bioinformatics*, 23:2686–2691, 2007.

[165] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J.R. Statist. Soc.*, B,63:411–423, 2001.

[166] M. Tischkowitz, J.-S. Brunet, L.R. Bégin, M.C.U. Huntsman, D.G.and Cheang, L.A. Akslen, T.O. Nielsen, and W.D. Foulkes. Use of immunohistochemical markers can refine prognosis in triple negative breast cancer. *BMC Cancer*, 7:134, 2007.

[167] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[168] R.R. Turner, D.W. Ollila, D.L. Krasne, and A.E. Giuliano. Histopathologic validation of the sentinel lymph node hypothesis for breast cancer. *Annals of Surgery*, 226:271–278, 1997.

[169] M. Van de Rijn, C.M. Perou, R. Tibshirani, P. Haas, O. Kallioniemi, J. Kononen, J. Torhorst, G. Sauter, M. Zuber, O.R. Köchli, F. Mross, H. Dieterich, R. Seitz, D. Ross, D. Botstein, and P. Brown. Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *Am J Pathol*, 161:1991–1996, 2002.

[170] L.J. Van't Veer, H.Y. Dai, M.J. van de Vijver, Y.D.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[171] P.F. Velleman and D.C. Hoaglin. *Applications, Basics and Computing of Exploratory Data Analysis*. Boston, Mass.: Duxbury Press, 1981.

[172] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. New York: Springer, 4th edition, 2002.

[173] M. Vuk and T. Curk. Roc curve, lift chart and calibration plot. *Metodološki zvezki*, 3(1):89–108, 2006.

[174] A.F. Wahl, K.L. Donaldson, and C. Fairchild. Loss of normal p53 function confers sensitization to Taxol by increasing G2/M arrest and apoptosis. *Nature Med*, 2:72–79, 1996.

[175] X.Y. Wang. *Fuzzy Clustering in the Analysis of Fourier Transform Infrared Spectra for Cancer Diagnosis*. PhD thesis, School of Computer Science, University of Nottingham, UK, 2006.

[176] X.Y. Wang and J.M. Garibaldi. A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis. In *Proceedings of second international conference in Computational Intelligence in Medicine and Healthcare*, pages 250–256, 2005.

[177] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[178] Cancer Research UK: Cancer Help website. CT scan, 2008. `http://www.cancerhelp.org.uk/help/default.asp?page=148`.

[179] Cancer Research UK: Cancer Help website. Breast cancer tests, 2009. `http://www.cancerhelp.org.uk/help/default.asp?page=3312`.

[180] Cancer Research UK website for Breast Cancer. Which treatment for breast cancer?, 2002. `http://www.cancerhelp.org.uk/help/default.asp?page=3318`.

[181] Cancer Research UK website for Breast Cancer. Cancerstats key facts on breast cancer, 2006. `http://info.cancerresearchuk.org/cancerstats/types/breast/?a=5441`.

[182] Cancer Research UK website for Breast Cancer. Breast cancer at a glance, 2007. `http://info.cancerresearchuk.org/cancerandresearch/cancers/breast/`.

[183] A. Weingessel, E. Dimitriadou, and S. Dolnicar. An examination of indexes for determining the number of clusters in binary data sets. Working Paper No.29, 1999.

[184] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, and J.R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, 98:11462–11467, 2001.

[185] C. Widakowich, E. de Azabuja, and T. Gil. Molecular targeted therapies in breast cancer: Where are we now? *Int J Biochem Cell Biol*, 39:1375–1387, 2007.

[186] Wikipedia. Hyperheuristic, November 2009. `http://en.wikipedia.org/wiki/Hyperheuristics`.

[187] M.P. Windham. Cluster validity for fuzzy clustering algorithms. *Fuzzy Sets and Systems*, 5:177–185, 1981.

[188] G.C. Wishart, M. Gaston, A.A. Poultsidis, and A.D. Purushotham. Hormone receptor status in primary breast cancer – time for a consensus? *Eur J Cancer*, 38(9):1201–1203, 2002.

[189] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[190] C.J. Witton, J.R. Reeves, J.J. Going, T.G. Cooke, and J.M.S. Bartlett. Expression of the HER1-4 family of receptor tyrosine kinases in breast cancer. *J. Pathol.*, 200(3):290–297, 2003.

[191] L.X. Xie and G. Beni. Validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.

[192] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[193] K.Y. Yeung and W.L. Ruzzo. An empirical study on principal component analysis for clustering gene expression data. Technical report, Department of Computer Science & Engineering, University of Washington, Seattle, US, 2000.

[194] K.Y. Yeung and W.L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[195] L.A. Zadeh. Fuzzy sets. *Inf. and Cont.*, 8:338–353, 1965.

[196] X. Zhang, C. Furtlehner, and M. Sebag. Frugal and online affinity propagation. In *Conférence Francophone sur l'Apprentissage (CAp' 2008), 29-31 May 2008, Ile de Porquerolles, France*, 2008.