An Exploration of Improvements to Semi-supervised Fuzzy c-Means Clustering for Real-World Biomedical Data

Daphne Teck Ching Lai
School of Computer Science
The University of Nottingham
Nottingham, United Kingdom

Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy
May 2014

To Grandma,
For your wonders,
inspirations
and love.

Declaration

The work in this thesis is based on research carried out at the Intelligent Modelling and Analysis Research Group, the School of Computer Science, the University of Nottingham, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Abstract

This thesis explores various detailed improvements to semi-supervised learning (using labelled data to guide clustering or classification of unlabelled data) with fuzzy c-means clustering (a 'soft' clustering technique which allows data patterns to be assigned to multiple clusters using membership values), with the primary aim of creating a semi-supervised fuzzy clustering algorithm that shows good performance on real-world data. Hence, there are two main objectives in this work. The first objective is to explore novel technical improvements to semi-supervised Fuzzy c-means (ssFCM) that can address the problem of initialisation sensitivity and can improve results. The second objective is to apply the developed algorithm on real biomedical data, such as the Nottingham Tenovus Breast Cancer (NTBC) dataset, to create an automatic methodology for identifying stable subgroups which have been previously elicited semi-manually.

Investigations were conducted into detailed improvements to the ssFCM algorithm framework, including a range of distance metrics, initialisation and feature selection techniques and scaling parameter values. These methodologies were tested on different data sources to demonstrate their generalisation properties. Evaluation results between methodologies were compared to determine suitable techniques on various University of California, Irvine (UCI) benchmark datasets. Results were promising, suggesting that initialisation techniques, feature selection and scaling parameter adjustment can increase ssFCM performance.

Based on these investigations, a novel ssFCM framework was developed, applied to the NTBC dataset, and various statistical and biological evaluations were conducted. This demonstrated highly significant improvement in agreement with previous classifications, with solutions that are biologically useful and clinically relevant in comparison with Sorias study [141]. On comparison with the latest NTBC study by Green *et al.* [63], similar clinical results have been observed, confirming stability of the subgroups.

Two main contributions to knowledge have been made in this work. Firstly, the ssFCM framework has been improved through various technical refinements, which may be used together or separately. Secondly, the NTBC dataset has been successfully automatically clustered (in a single algorithm) into clinical sub-groups which had previously been elucidated semi-manually. While results are very promising, it is important to note that fully, detailed validation of the framework has only been carried out on the NTBC dataset, and so there is limit on the general conclusions that may be drawn. Future studies include applying the framework on other biomedical datasets and applying distance metric learning into ssFCM.

In conclusion, an enhanced ssFCM framework has been proposed, and has been demonstrated to have highly significant improved accuracy on the NTBC dataset.

Acknowledgements

This work would not have materialised without the guidance and support of my supervisor, Prof. Jonathan M. Garibaldi, whom I am thankful for. I also thank him for the funding of all conferences I attended. To Dr. Daniele Soria, I thank him for the insightful discussions on breast cancer classification and for the advice given which has provided me confidence.

To Dr. Grazziela Figueredo, I thank her for being a tremendous help with my thesis. For the support and assistance and for always bringing laughter along, I thank Dr. Christopher M. Roadknight. For encouraging me to apply for a job at the Universiti Brunei Darussalam, which has led me towards a PhD path, I thank Dr. Pg. Hj. Nor Jaidi Pg. Hj. Tuah.

To my examiners, Prof. Uwe Aickelin and Prof. Azzam F. G. Taktak, I am thankful for their invaluable advice which has improved my thesis.

I thank Universiti Brunei Darussalam for recommending me to pursue a PhD and the Government of Brunei Darussalam for funding my studies.

To the amazing members of The Intelligent Modelling & Analysis Research Group (past and present), I thank them for making my studies at the University of Nottingham so fun, productive and fulfilling. I also thank Rosy, Ros, Sein, Arif, Elvy, Hong and Zimah for maintaining a close and supportive rapport despite our distances. Jo, Hui, Jon, Dan, Chan, Irene, Puff and Darlin, thank you for your constant love and support. To my family, I am thankful for their love, prayers and encouragement. I also thank my partner for being my unfailing pillar of support.

Contents

List of Figures

List of Tables

List of Abbreviations

κ	Cohen's Kappa Index
ART	Adaptive Resonance Theory
BIC	Bayesian Information Criterion
CE	Cluster Estimation
CFS	Correlation-based Feature Selection
CKM	Consensus K-means
CSQ	Chi Square
CV	Cross Validation
DML	Distance Metric Learning
EKKZ	Euclidean ssFCM with KKZ initialisation
EKKZ30	Euclidean ssFCM with $\alpha = 30$ and KKZ initialisation
EM	Expectation Maximisation
ENB	Euclidean ssFCM with 15 important features identified using NB-RFE and ssFCM feature selection
ENB30	Euclidean ssFCM with $\alpha = 30$ and 15 important features identified using NB-RFE and ssFCM feature selection
FCM	Fuzzy C-means
GLMNET	Generalised Linear Models with Elastic Nets
GR	Gain Ratio
HC	Hierarchical Clustering
HDDA	High Dimensional Discriminant Analysis
IG	Information Gain
KKZ	Initialisation technique by Katsavounidis, Kuo and Zhang
KM	K-means
KNN	k-Nearest Neighbours
KSVM	Kernel Support Vector Machines
LDA	Linear Discriminant Analysis
LVQ	Learning Vector Quantisation
MAP	Maximum *A Posteriori*
MBIC	Model-based clustering via Bayesian Information Criterion
MDA	Mixture Discriminant Analysis
MLE	Maximum Likelihood Estimation

n.c	Not classified
NB	Naive Bayes
NB-RFE	Naive Bayes-Recursive Feature elimination
NMI	Normalised Mutual Index
NNET	Neural Networks
NPI	Nottingham Prognostic Index
NTBC	Nottingham Tenovus Breast Cancer dataset
OneR	One Rule Classification
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
PCCA	Pairwise-constrained Competitive Agglomeration
PID	Pima Indians Diabetes
RF	Random Forests
RF-RFE	Random Forest-Recursive Feature Elimination
SCS	Simple Cluster Seeking
ssFCM	Semi-supervised Fuzzy c-Means
SU	Symmetrical Uncertainty
SVM-RFE	Support Vector Machine-Recursive Feature Elimination
TSVM	Transductive Support Vector Machines
WOBC	Wisconsin Original Breast Cancer

1 Introduction

This thesis describes the application of semi-supervised Fuzzy c-Means (ssFCM) algorithm on real-world biomedical data to assist clinicians in decision making. Investigations in distance metric, initialisation, feature selection and scaling parameter adjustment for improvement to ssFCM are explored and tested on other popular datasets. Two applications of ssFCM are demonstrated. First, a novel integrated ssFCM framework which incorporates other machine learning techniques, initialisation and feature selection is applied on the NTBC dataset to assign new patients into the six subgroups identified by Soria *et al.* [141] and to demonstrate how the classification results are used to assist clinicians. Secondly, the integrated ssFCM framework is applied to identify stable and clinically meaningful subgroups in the data set. In this chapter, the background and motivation behind this research are described and its aims and objectives are highlighted. The organisation of this thesis is also outlined.

1.1 Background and motivation

Clustering is a popular exploratory tool for identifying groups of interests in biomedical data [46, 154, 3, 141]. It is a form of unsupervised learning which organises unlabelled patterns of a data set into groups called clusters based on their similarities. Classification, on the other hand, is supervised because labelled patterns are used to train (teach) the algorithm. Once trained, the algorithm can assign unlabelled patterns into categories called classes, based on the model learned during training. While labelled patterns provide more information than unlabelled ones, they take time to collect and are often scarce. To overcome this, researchers have been developing

semi-supervised learning techniques [30, 22, 166]. These techniques use available labelled patterns to provide additional information to aid clustering or classification of unlabelled patterns. In a semi-supervised clustering environment, the available class information from labelled patterns provide additional structural information about the data, which makes simultaneous class assignment of unlabelled patterns possible. Thus, classification can be achieved using semi-supervised clustering techniques [121].

Many semi-supervised clustering techniques have been proposed such as those summarised in [84], but it is not feasible to investigate them all. Instead, the study on semi-supervised Fuzzy c-Means (ssFCM) is focused in this thesis. ssFCM involves the use of Fuzzy set theory where known membership values of labelled patterns help organise unlabelled patterns according to their similarities. Memberships between values of zero and one are used to indicate the degree of "belongingness" a pattern has to clusters. Not only does this form of representation gives a more realistic model as compared to binary-valued representation, it also allows adjustment between generality and precision of pattern recognition [123]. This feature of Fuzzy c-Means is favoured in clustering or classification of biomedical data as there are often no distinctive boundaries separating the classes.

Much of the developments in ssFCM have been extended from the popular Fuzzy c-Means (FCM) algorithm due to its clustering capabilities which can be enhanced by its flexibility to work with different techniques at different levels (objective function [103, 64, 47], cluster centre [15, 165] or membership [21]). Fuzzy clustering with objective function such as FCM is dynamic allowing patterns to move from one cluster to another to minimize the objective function and knowledge about the shape or size of clusters can be incorporated in the objective function using appropriate distance measure [56]. However, FCM faces several problems, which are also common to ssFCM. The main problems are the unknown number of clusters, sensi-

tivity to initialisation and sensitivity to noise. In a dataset where no labels exist, the number of useful clusters is unknown. To overcome the unknown number of clusters problem, several initial cluster numbers are tried and the produced partitions are then evaluated using several cluster validity indices such as partition coefficient, partition entropy coefficient [16], and Xie-Beni index [161]. The optimal validity values indicate the correct number of clusters. While this has been a popular approach to find the number of clusters, the approach has several problems [70], such as monotonous dependency on the number of clusters. Sensitivity to initialisation refers to the sensitivity of the algorithm to initial parameters such as the initial membership values assigned to patterns or the initial cluster centres. The sensitivity to noise refers to the sensitiveness of the algorithm to patterns which may not reflect the true nature of the dataset. In ssFCM, some of these problems remain unresolved in addition to other issues such as the representation of high-dimensional data, balance of between supervised and unsupervised training and equal population of clusters.

Many approaches involving the modification of objective function have been employed to improve semi-supervised Fuzzy c-means clustering. This can make the algorithm complex to use as they contain many user-defined parameters. Our approach, however, focuses on integrating existing initialisation and feature selection techniques into an existing, simple-to-implement semi-supervised Fuzzy c-means framework with investigation in distance metric and scaling parameter adjustment α. While there are many existing initialisation and feature selection techniques available, to the best of our knowledge, no such study to improve ssFCM has been conducted and this work aims to fill this research gap. Furthermore, investigation to determine the most suitable distance metric is not usually done even though distance metric is widely known to be important in data representation for clustering. In addition, the adjustment of scaling parameter adjustment

α has been reported to affect clustering [23] but, users of ssFCMs do not always exploit this feature in ssFCM. Thus, the motivation for our work is to demonstrate that ssFCM accuracy can be improved using a simple-to-implement ssFCM with investigation in distance metric, initialisation, feature selection and adjustment in scaling parameter.

Cancer is a leading cause of death worldwide, with breast cancer being the most common in women [51, 26, 139]. In 2008, more than 1.38 million women were diagnosed [51, 26, 139]. The severity of this disease brings urgency to provide decision making support for clinicians to achieve more accurate prognosis [154] and to administer specialised treatment for patients based on the breast cancer subtypes (subgroups or classes) [1]. Prognosis is the prediction of the likely outcome of one's current medical condition such as survival rate or survival time. With more accurate prognosis, predictions can to be made to patients belonging to similar subgroups with possible risks of recurrence or death.

Due to advances in machine learning techniques [85, 75] and in biomedical technologies for data collection, researchers are now applying machine learning techniques to interpret complex biomedical data and to identify important patterns in the data such as their subgroups that can give new insights to the disease, in the hope to increase accuracy in prognosis. Recent studies using high-throughput molecular technologies and machine learning methodologies are showing evidence of biological differences in breast cancer subgroups and clinical relevance of these subgroups to survival outcome [3, 141]. The task of identifying important subgroups is a complex one as there are no standard correct answers to compare with. Instead, the features (factors, biomarkers or genes) that characterise these subgroups must be consistent with biological findings from existing literature subsequently indicating the biologically meaningfulness of these subgroups, and that they must be clinically relevant [9].

Six subgroups of breast cancer were found by applying hierarchical clustering to the Nottingham Tenovus Breast Cancer (NTBC) immunohistochemical dataset. The sixth group, however, contained only four patients [3]. A subgroup with only four patients in a dataset of 1076 patients may not be considered useful and may not hold sufficient evidence of its actual existence. To identify breast cancer subgroups and address their stability in NTBC, Soria *et al.* [141] used a consensus clustering methodology where a consensus is reached from several different clustering solutions. They identified six novel subgroups of breast cancer (Soria's classification for short) and determined the key biomarkers that characterise these subgroups. These subgroups are regarded as clinically useful as they are not only biologically meaningful, but are also relevant to clinical data. However, the methodology used was semi-manual, involving visual inspection and the use of heuristics and other techniques to aggregate results from different unsupervised clustering techniques. As no single unsupervised clustering technique has been found to do this so far, the development of a fully automated method (post-initialisation) for identifying these same six subgroups is necessary.

1.2 Aims and objectives

The main aim of this research is to develop novel ssFCM methodologies that are applicable to real biomedical data such as the NTBC while addressing some of the ssFCM problems discussed. To fulfill the main aim, the following main and sub-objectives are to be achieved:

1. Develop a novel ssFCM algorithm that can address the problem of initialisation sensitivity and can improve clustering or classification results. The sub-objectives are to:

   (a) Investigate in the performance of different semi-supervised Fuzzy

c-Means clustering algorithms such as distance-based ones by Pedrycz and Waletzky [121], Zhang *et al.* [165], Li *et al.* [101] and Endo *et al.* [47].

(b) Ensure robustness in the algorithms using various evaluation techniques such as accuracy rate, Kappa's Cohen Index, Normalised Mutual Index and cross-validation.

(c) Investigate in the different initialisation techniques that can help improve clustering or classification in ssFCM.

(d) Explore other techniques with ssFCM from those shown to be successful in Fuzzy clustering or in machine learning algorithms such as feature selection and distance metric learning.

(e) Apply the modified ssFCM algorithms on different data sets to demonstrate generalisation of the algorithms to other datasets such as popular UCI [54] datasets.

2. Apply the developed clustering algorithm on a real biomedical data, such as breast cancer data. Ideally, the developed algorithm is an automatic methodology for identifying the same six breast cancer subgroups identified by Soria *et al.* [141] in the NTBC dataset, as part of a confirmatory study to address the stability of the six subgroups. To the best of our knowledge, no one clustering algorithm has been applied on breast cancer immunohistochemical data for classification into distinct biological subgroups. The sub-objectives are to:

(a) Investigate in the accuracy of ssFCM methodologies in the clustering and classification of the NTBC dataset.

(b) Compare ssFCM performance with other well-known classification techniques to determine its performance in comparison with other techniques.

(c) Solve real-world problems by building a framework that incorporate investigated approaches from the first objective in a working system which can help provide support to clinicians in decision-making based on the clustering or classification results.

(d) Address the issue of subgroup stability by consistently reproducing the six subgroups using ssFCM and comparing agreement with other clustering techniques, which in turns validate the subgroups.

This research work is 1) an exploratory study from the technical point of view with investigations into the application of ssFCM and other machine learning techniques for classification and 2) a confirmatory study from a clinical point of view with identification of the same six subgroups found by Soria *et al.* [141] to demonstrate stability of these six subgroups.

In this work, however, the problems of unknown number of clusters and noise sensitivity are not addressed. Soria *et al.* [141] had identified the optimal cluster number to be six using cluster validity indices on K-means and Partitioning Around Medoids clustering solutions. The current NTBC dataset of 1076 patients is considered informative by clinicians [3] and has been reduced from the original 1944 patients. Thus, any patients that are not representative to the population have been assumed to be removed.

1.3 Organisation of the thesis

The organisation of this thesis is as follows. In Chapter 2, a literature review touching on clustering techniques, ssFCM and breast cancer classification is presented. Initialisation and Evaluation techniques are also reviewed. The motivation for the focus in ssFCM is explained, including its developments and applications. As the research is focused in the development of ssFCM for application on a real biomedical dataset, it is

vital to understand the developments in the identification of breast cancer subgroups in the Nottingham Tenovus Breast Cancer dataset as well as other similar dataset and their challenges, given there are no definite correct subgroups to compare with.

In Chapter 3, three preliminary investigations are reported. The first is a comparative study in distance-based ssFCM algorithms, the second is the application of ssFCM for breast cancer classification with exploration in distance metrics and lastly, a comparison of ssFCM classification with other classifiers. As there are many existing ssFCM algorithms, several algorithms shown to produce good classification results are chosen and compared on popular benchmark datasets. The ssFCM deemed to be best performing (in terms of average accuracy) is applied for breast cancer classification. Further comparisons with other classifiers are conducted to show its suitability for breast cancer classification. In this chapter, sub-objectives 1a, 1b, 2a and 2b are fulfilled.

Three approaches for improvement of ssFCM classification are reported in Chapter 4; initialisation techniques, feature selection and adjustment of ssFCM's scaling parameter α. Initialisation techniques are employed to overcome the problem of initialisation sensitiveness in ssFCM. Feature selection techniques are employed to identify important features and reduce the number of features used for classification. Thus, the time and cost of running clinical tests (or procedures) for collecting measurements of these features are reduced. There has been no definitive guide as to the best setting for α in ssFCM and thus, its choice is experimental. In these studies, the positive experimental results based on applications on UCI datasets as well as the NTBC dataset are presented. Sub-objectives 1c, 1d, 1e and 2a are fulfilled in this chapter.

An integrated framework, using initialisation and feature selection, to predict class labels of new or unlabelled breast cancer patients, for assisting

clinicians in decision making is presented in Chapter 5. This framework is based on the findings detailed in Chapters 3 and 4 and fulfill the second objective of this research work. The strategy behind this integrated framework is explained and experimental results that demonstrate the framework are presented. Sub-objectives 2c and 2d are fulfilled in this chapter.

Based on previous investigations in approaches of improvement for ssFCM, a ssFCM methodology which incorporates these approaches, specifically nitialisation technique by Katsavounidis, Kuo and Zhang (KKZ) [90] and adjustment of α, in Chapter 6 is applied to identify six stable breast cancer subgroups. An investigation using two different reduced panels of biomarkers, one panel identified in Chapter 4.2 and one identified in [140] is conducted. The stability of the resulting subgroups are evaluated based on agreement with subgroups identified by unsupervised clustering algorithms Consensus K-means (CKM) and Model-based clustering via BIC (MBIC). The stability of the subgroups from the different clustering solutions can help ascertain which panel of biomarkers are relevant. Furthermore, the stability of seven breast cancer subgroups (with the HER2 group split into two) is investigated by comparing results based on clinical evaluation with the latest breast cancer subgroups identified in [63]. Sub-objectives 2c and 2d are fulfilled in this chapter.

In Chapter 7, this research work is concluded with a summary of results including the main contributions, limitations and future work. In addition, peer-reviewed, accepted papers and oral presentations derived from this work are listed.

1.4 Contribution to knowledge

This research work has led to the development of a novel integrated ssFCM-based framework with the incorporation of approaches such as initialisation, feature selection and/or adjustment of scaling parameter. These ap-

proaches of improvement have been individually investigated with ssFCM on other different UCI datasets and have shown to increase ssFCM accuracy. The framework has been applied in two ways; firstly, the ssFCM framework is applied to perform a classification task of assigning class labels to new or unlabelled patients and secondly, the framework performs a clustering task of identifying clinically useful and stable breast cancer subgroups with full retention of Soria's classification using a reduced panel of biomarkers. To evaluate whether the subgroups found are clinically useful,the biomarker profiles of each subgroup are compared with those by Soria *et al.* [141] for biological meaningfulness and conduct statistical analysis to measure clinical association. The subgroups identified demonstrated stability, showing high agreement with other clustering algorithms. In analysing clinical association of the identified subgroups, clinical data can be stratified to allow clinicians to identify relevant clinical characteristics (including survival outcome) which are associated with these subgroups. Therefore, it is hoped such information interpreted through visual and statistical analysis based on the identified subgroups can assist clinicians in predicting survival outcome of patients and in planning of treatments.

Thus, two main contributions can be drawn out of this work. Firstly, the novel application of initialisation in ssFCM, which has shown promising results to increase ssFCM results, in UCI datasets as well as in the NTBC dataset. Secondly, the development of a novel integrated ssFCM-based framework to classify new or unlabelled patients into the same six breast cancer subgroups as Soria *et al.* [141]. Experimental findings show that accuracy increased significantly using the ssFCM framework as opposed to an unsupervised approach. By classifying new or unlabelled patients,it is hoped that a more accurate model have been provided for the prediction of breast cancer types for future patients. The framework is also applied for the identification of stable breast cancer subgroups using the integrated

framework, where no one single clustering algorithm has been able to do so far. On comparison of clustering results with those obtained using CKM and MBIC, higher agreement that those in [141] were found, indicating higher stability in these subgroups. The second main contribution can be used by clinicians as decision making support in two application areas.

Based on this research work, several refereed papers have been produced two journal papers and six conference papers. A detailed list of these publications is shown in Chapter 7.3.

- This page is empty -

2 Literature Review

One major contribution of this work is the application of semi-supervised Fuzzy c-Means on a real, biomedical dataset to assist clinicians in decision support. Two application areas are focused on; 1) the automatic classification of new breast cancer patients in the NTBC dataset with high accuracy to Soria's biological classification of breast cancer [141] and 2) the identification of stable breast cancer classes. Due to the focus in ssFCM, the motivation for the development of ssFCM algorithm as an alternative to predecessor clustering algorithms is reviewed. In addition, a study on ssFCM latest developments is presented. Some developments for tackling problems in clustering approaches relevant to this research, particularly with regards to initialisation (for centroid-based clustering algorithms, such as ssFCM) is also reviewed. Other potential approaches of improvement such as feature selection is also reviewed. As the aim is towards real medical applications, relevant background work in the use of clustering techniques for breast cancer classification and identify the issues faced by existing methodologies are covered. In this way, it is hoped that the knowledge gaps which this research aim to fill are identified.

2.1 Definitions

As this research is multidisciplinary covering data mining and biomedical areas, the following terminologies are used:

Definition 1 (Data pattern.) A data pattern (data instance, feature vector, object, observation or datum) $\mathbf{x_j}$ is a vector representing data point j in a data matrix $\mathbf{X}$. The data matrix is of size $N \times n$, containing data patterns $\mathbf{x_1}, ..., \mathbf{x_N}$ with n features.

Definition 2 (Features.) Features (attributes, dimensions or biomarkers) are measurements of properties which describe data patterns. Each data pattern has n number of features such that $\mathbf{x_j} = \{x_{j1}, ...x_{jn}\}$. Hence, n is the dimensionality of the data matrix or pattern.

Definition 3 (Labelled data.) Labelled data (examples or training data) are data patterns that have prior knowledge of the clusters or classes they belong to, while unlabelled data do not have such prior information. They are referred to as $\mathbf{x_j^l}$ and their labels (or classes) are referred as y_i.

Definition 4 (Unsupervised learning.) Unsupervised learning involves learning from data without labelled data. The data patterns are unlabelled, that is, no prior knowledge on which clusters the data patterns belong to is known. Clustering is a form of unsupervised learning. Further explanation of clustering is found in Chapter 2.3.

Definition 5 (Supervised learning.) Supervised learning uses examples, which have both inputs and known outputs, to learn a function in order to map new examples to outputs. These examples are labelled data with prior information such as cluster or class labels. Classification is a form of supervised learning and is further explained in Chapter 2.2.

2.2 Pattern recognition

In machine learning, pattern recognition is the task of assigning labels to data patterns by learning from data [17, 43]. There are several different types of labels; real-value, categorical, sequenced and so forth [91]. For assigning categorical labels to data patterns, two learning approaches can be used; supervised (classification) and unsupervised (clustering) [43]. A more recent approach is semi-supervised learning [30, 166].

Classification is a form of supervised learning which assigns data patterns into known, meaningful categories called classes [43]. To do this,

classification algorithms learn the mapping between the data patterns and their classes from a training set made up of $(\mathbf{x_j^l}, y_i)$. The mapping f forms a model for the different classes based on the values of predictor features such that $f : X \leftarrow Y$. Once this mapping is accurately learned, it can be applied to new, unlabelled data patterns, also called a test set, to determine their classes. The training set are made up of labelled data which have prior knowledge of the data patterns' classes (also known as class labels).

Labelled data are often scarce because they are time-consuming and labour-intensive to collect. Due to their limited availability, they are often not enough of them to learn an accurate mapping from and thus, to assign data patterns accurately to their respective classes. To learn an accurate mapping and discriminate between classes, there has to be enough training data to represent the variability of feature values for data patterns in the same class relative to the differences between feature values for data patterns in different classes [43]. Often in biomedical datasets such as breast cancer data, the clinicians do not know exactly how many subtypes of breast cancer there are and it is hard to determine the number of useful subgroups possibly exist and which labels to associate the data with.

In situations where labels are few or unknown, clustering techniques can be used where data patterns that are most similar are grouped together and the groups validated. Subsequently, meaningful class labels are manually assigned. Thus, clustering produces initial categories [70]. However, clustering is a more challenging problem than classification as the clusters that are of interest may not be easily extracted or that the discovered clusters may not be valid or meaningful [84]. Semi-supervised clustering techniques allow 1) the use of a few labels to guide the identification of clusters that are of interest and 2) the determination of classes for unlabelled or future data patterns simultaneously.

The difference between classification and clustering can be confusing.

Tan *et al.* [147] explained that clustering, when used for understanding data, is the partitioning of data into groups, while classification is the assigning of data patterns to these groups based on a learned model. These definitions are also used in machine learning. Tan *et al.* further explained that clustering is a form of classification as it generates class (or cluster) labels for data patterns from the data itself. The clusters discovered are potential classes. In some fields such as biology and ecology, classification refers to cluster analysis, a form of unsupervised learning (unsupervised classification), which reflects Tan's second definition for clustering as generators of class labels. A different distinction between clustering and classification is explained by Jain [84] such that these learning approaches use unlabelled data (unsupervised) and labelled data (supervised) respectively. Jain's explanation are in line with the definitions used in machine learning as labelled data are used in classification to learn a model.

The boundary between clustering and classification becomes a blur in semi-supervised clustering algorithms. It is debatable whether semi-supervised clustering algorithms is capable of directly performing classification tasks (in machine learning terms) as there may be confusion between the definition of clusters (geometrical structure) and classes (logical structure) of the data. This may not be a problem when a class is represented by a cluster as demonstrated in several literatures such as by Pedrycz and Waletzky [121] and by Li *et al.* [101], but classification task using a clustering algorithm becomes complicated when a class is represented by several clusters as additional mechanism would be required to align these clusters to a class [23], to reflect the classification and not the clustering. ssFCM can be a classification or clustering technique depending on the intention of application, which can either be to predict the class labels of new data pattern or to identify meaningful clusters.

In this research, ssFCM is employed as a classification technique to

assign new data patterns into known groups using a learned model (classification in machine learning) and as a clustering technique with the use of labelled data to guide the identification of meaningful clusters (clustering in machine learning for biological classification). In a classification environment, the accuracy of ssFCM in assigning the new data patterns (unlabelled) in the correct classes (known groups) is of interest. Whereas, in a clustering environment, the subgroups formed by using all the data patterns (both labelled and unlabelled) are of interest. Both the classification and clustering tasks will be discussed in greater detail in Chapters 3, 4, 5 and 6 based on the investigations carried out.

These learning approaches have been popularly applied in many fields such as ecology [48], logistics [146], economics [41] and medicine [46]. In medical areas, these approaches are used to identify subgroups of a disease in order to help predict survival outcomes and provide support in decision making in the choice of specialised treatment for different subgroups [9].

2.3 Clustering

Clustering is an unsupervised learning approach. Clustering involves the grouping of similar (unlabelled) data patterns that appear to form natural clusters together. Mackay [105] highlighted the following motivations for clustering. Clustering has predictive power allowing one to predict those data patterns that share the same cluster will have similar properties. A summary of clusters can be communicated using cluster centres, which are representatives of the clusters. In addition, the interesting data patterns that deserve further attention can be highlighted when there is a failure to build a good cluster model. Furthermore, the competitive learning in clustering means that clusters compete to own data patterns.

In data mining, clustering is a popular exploratory tool to find structural information hidden within the data that can identify groups of inter-

est [121, 159]. It has also been used to perform classification tasks where the discovered clusters are validated for meaningfulness or usefulness and labelled as classes. In this section, Hierarchical Clustering, K-means, Fuzzy c-Means and model-based clustering via Bayesian information criterion, which are applied in this work, are introduced.

2.3.1 Hierarchical clustering

Based on the similarity of data patterns, hierarchical clustering (HC) [50, 43] iteratively merges clusters, forming a hierarchy of clusters known as a dendrogram. This type of HC is called agglomerative hierarchical clustering and takes a bottom-up approach, with each data pattern initially located in one cluster. The similarity of data patterns is determined using a distance metric, $d()$ (such as Manhattan, Euclidean and Mahalanobis) [43] and a linkage criterion of data patterns. Distance metrics are covered in further details in Chapter 2.3.5. The algorithm is defined as follows:

1. Initialise with each data pattern in its own cluster.
2. Find the nearest cluster pair **A**, **B** such as $d(\mathbf{A},\mathbf{B})$ is minimised.
3. Merge **A** and **B** to form a new cluster.
4. Repeat step (2) and (3) until the desired number of clusters or a $d()$ threshold is achieved.

Table 2.1 shows some of the linkage criteria used in hierarchical clustering:

Table 2.1: Linkage criteria in Hierarchical clustering.

Linkage criteria	Formula
Complete linkage	$\max\{d(a,b) : a \in A, b \in B\}$
Single linkage	$\min\{d(a,b) : a \in A, b \in B\}$
Average linkage	$\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} d(a,b)$

Other linkage criteria include the sum of all intra-cluster variance and Ward's criterion which uses minimum variance [50]. The divisive hierarchical clustering takes a top-down approach and splits clusters instead. The dendrogram reveals clusters, subclusters and subsubclusters at a much greater detail. Once clusters have been merged or split, the clusters where the data patterns belong to are locally determined and cannot be changed unless in accordance with the path of the dendrogram in preceding iterations [8]. This raises two issues; 1) the merge or split is final and 2) the decision to merge or split clusters is based on local information of clusters, rather than global information. This means, for instance, data patterns assigned to one parent cluster will not be reassigned to any clusters belonging to another parent cluster in the divisive approach.

2.3.2 K-means

The K-means (KM) [106] is a centroid-based clustering technique which partitions N data patterns into k clusters according to the closeness of data pattern x_j to cluster centre (or centroid) v_i. A cluster centre is determined based on the average of all the data patterns in the cluster. In K-means, data patterns either belong or do not belong to a cluster. Thus, the belongingness of a data pattern to a cluster is binary and its assignment is known as hard assignment. The aim of K-means is to minimize the intra-cluster squared distances given in (2.1) below:

$$J(V) = \sum_{i=1}^{k} \sum_{j=1}^{N_i} ||\mathbf{x_j} - \mathbf{v_i}||^2 \tag{2.1}$$

where N_i is the number of data patterns belonging to cluster i, k is the number of clusters and $||\mathbf{x_j} - \mathbf{v_i}||$ is the Euclidean distance between data pattern $\mathbf{x_j}$ to cluster centre $\mathbf{v_i}$.

The algorithm is described in [105] as an iteration of two steps listed

below, given some form of initialisation, such as a random initialisation of cluster centres. The algorithm terminates when $J(V)$ or $\mathbf{v_i}$ stabilise:

1. Assignment step: a data pattern is given an assignment c_j to the closest cluster centre (mean) as follows:

$$c_j = \arg\min_i (d(\mathbf{x_j}, \mathbf{v_i}), \forall i \in k) \tag{2.2}$$

2. Update step: the cluster centres are updated as follows until they stabilise:

$$\mathbf{v_i} = \frac{1}{|i|} \sum_{\mathbf{x_j} \in i} \mathbf{x_j} \tag{2.3}$$

Unlike hierarchical clustering, K-means can assign a data pattern to a different cluster, where it previously belonged to or not, at different iterations depending on the closeness to the cluster centres. Choosing initial cluster centres for K-means is challenging as the common practice of randomly selected initial centres leads to poor clustering solutions. One approach is to first cluster a small sample using hierarchical clustering and the cluster centres generated are subsequently used to initialise K-means [147]. Another issue of K-means is the number of clusters requirement, which is often unknown. A range of cluster validity techniques [158, 119, 19, 110] are available to identify the optimal number of clusters. As cluster validity in terms of number of clusters is out of the scope of the research aims, it will not be further discussed. Due to K-means locality-based (using centroids) and unstructured nature, it is not suitable for clustering non-globular clusters, or clusters of different sizes and densities, and have difficulty with data that contain outliers [8, 147].

2.3.3 Fuzzy c-Means

An extension of the K-means is the Fuzzy c-Means (FCM) algorithm. FCM was first proposed by Dunn [44] and was then improved by Bezdek [17]. Unlike the hard assignment in K-means, data patterns can be assigned to multiple clusters using membership values (soft assignment) [17, 43]. The aim of FCM is to minimise the objective function (2.4) so that data patterns similar in structure are assigned the same cluster.

The objective function is defined as:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^{p} d_{ij}^{2} \tag{2.4}$$

where

- $1 < p < \infty$ is a fuzzifier parameter,
- u_{ij} is the membership value of data pattern j in cluster i with values between 0 and 1 and
- d_{ij} denotes similarity between data pattern j and cluster i that can be calculated using a distance metric d_{ij}.

The algorithm involves iteratively calculating the cluster centres and the partition matrix to minimise the objective function until a termination criterion is satisfied. It is summarised as follows:

1. Initialise partition matrix $\mathbf{U} = [u_{ij}]$, $\mathbf{U}^{(0)}$.
2. Calculate cluster centres $\mathbf{V} = [\mathbf{v_i}]$ with $\mathbf{U}$ using equation:
$$\mathbf{v_i} = \frac{\sum_{j=1}^{N} u_{ij}^{n} \mathbf{x_j}}{\sum_{j=1}^{N} u_{ij}^{m}} \tag{2.5}$$
3. Update partition matrix, U using the following equation:
$$u_{gj} = \frac{1}{\sum_{i=1}^{c} \left(\frac{||\mathbf{x_j} - \mathbf{v_g}||}{||\mathbf{x_j} - \mathbf{v_i}||} \right)^{\frac{2}{m-1}}} \tag{2.6}$$
4. If $||\mathbf{U'} - \mathbf{U}|| < \epsilon$, stop. Else, go to step 2.

Like K-means, FCM also suffer from the number of clusters requirement and sensitivity to initialisation. For handling non-globular clusters, researchers have devised ways to manipulate the distance metrics to adapt to the different shapes of clusters, such as fuzzy c-rectangular shells algorithm for the detection of rectangles [81].

2.3.4 Model-based clustering via Bayesian information criterion

Fraley and Raftery [53] implemented a model-based clustering (MBIC) which uses the Maximum *A Posteriori* (MAP) estimate from a Bayesian analysis to estimate model parameters, instead of Maximum Likelihood Estimation (MLE) in the Expectation-Maximization (EM) algorithm and a modified Bayesian Information Criterion (BIC) for model selection. MAP is used to avoid the failure of the EM algorithm in the presence of singularities or degeneracies. The mixture model with density for generating data $y = (y_1, ..., y_n)$ in model-based clustering is defined as:

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k), \tag{2.7}$$

where

- $f_k(y_i|\theta_k)$ is a probability distribution with parameters θ_k,
- τ_k is the probability of belonging to the k^{th} component,
- $\theta_k = (\mu_k, \Sigma_k)$, μ_k are the means and Σ_k the covariances of f_k.

These model parameters are estimated using MLE in the EM algorithm.

To eliminate EM failure to converge due to singularity in covariance estimate, the authors in [53] proposed a prior distribution on the parameters that can eliminate the singularity problem while maintaining stability on results obtainable without a prior probability . The Bayesian predictive

density for the data is in the form:

$$\mathcal{L}(Y|\tau_k, \mu_k, \Sigma_k)\mathcal{P}(\tau_k, \mu_k, \Sigma_k|\theta), \tag{2.8}$$

where $\mathcal{L}$ is the mixture likelihood:

$$\begin{aligned}\mathcal{L}(Y|\tau_k, \mu_k, \Sigma_k) &= \prod_{j=1}^{n}\sum_{k=1}^{G}\tau_k\phi(y_i|\mu_k, \Sigma_k) \\ &= \prod_{j=1}^{n}\sum_{k=1}^{G}\tau_k|2\pi\Sigma_k|^{-\frac{1}{2}} \\ &\quad \exp\left\{-\frac{1}{2}(y_j-\mu_k)^T\Sigma_k^{-1}(y_j-\mu_k)\right\},\end{aligned}$$

and $\mathcal{P}$ is a prior distribution on the parameters τ_k, μ_k, σ_k and θ. The f_k in (2.7) is the multivariate Gaussian density ϕ with parameters μ_k as its mean and Σ_k as its covariance.

The BIC [138] selects the best fitted model from a finite set of models using maximum likelihood. It is defined [53] as:

$$BIC \equiv 2\log\mathcal{L}_{max} - k\log(N) \tag{2.9}$$

where $\mathcal{L}_{max}$ is the maximum likelihood of the estimated model, k the number of parameters in the model and N the number of data patterns used in the estimation. The BIC is modified by replacing the first term in (2.9), $2\log\mathcal{L}_{max}$ by twice the log-likelihood evaluated using MAP in (2.8).

2.3.5 Distance metrics

Similarity between data patterns is important in defining a cluster. Distance metrics are used to measure similarity of patterns and clusters. The performance of a clustering algorithm can be greatly affected by how accurately the selected distance metric represents similarity.

Euclidean distance

The Euclidean distance metric [43] forms spherical clusters and does not reflect scale differences among dimensions in high-dimensional datasets, that is, it is not scale invariant. Dimensions with smaller scales will have less influence on the distance than dimensions with larger scales [85]. With a high number of dimensions having different scales, it can exponentially affect the Euclidean distance, which in turn can negatively affect clustering results. The Euclidean distance is computed as follows:

$$d_E^2(i,j) = ||\mathbf{x_j} - \mathbf{v_i}||^2 \tag{2.10}$$

where $\mathbf{x_j}$ is a data pattern and $\mathbf{v_i}$ is a cluster centre.

Mahalanobis distance

The Mahalanobis distance is formally defined [107, 43] as:

$$d_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T \mathbf{S^{-1}} (\mathbf{x} - \mu)} \tag{2.11}$$

It is the distance between a vector $\mathbf{x} = (x_1, x_2, ...x_n)^T$ which belongs to a group of vectors with mean $\mu = (\mu_1, \mu_2, ...\mu_n)^T$ and covariance matrix $\mathbf{S}$ of the group. Due to the covariance matrix, it identifies ellipsoidal clusters. It is scale-invariant as it takes into account of the correlation within the data. The Euclidean distance is a special case of Mahalanobis distance where the covariance matrix is an identity matrix [163].

Gustafson and Kessel [66] introduced the Fuzzy covariance matrix (2.14) to generalise the distance metric in FCM's Mahalanobis distance. This was to represent patterns in a more natural manner as Fuzzy weights are used to make the clusters more adaptive. This Mahalanobis distance is computed as follows:

$$d_{fM}^2(i,j) = (\mathbf{x_j} - \mathbf{v_i})^T \mathbf{M_i} (\mathbf{x_j} - \mathbf{v_i}) \tag{2.12}$$

where $\mathbf{M_i}$ is a positive definite matrix, its inverse defined as:

$$\mathbf{M_i}^{-1} = \left[\frac{1}{\rho_i det(\mathbf{P_i})}\right]^{\frac{1}{n}} \mathbf{P_i} \tag{2.13}$$

and $\mathbf{P_i}$ is a Fuzzy covariance matrix defined as:

$$\mathbf{P_i} = \frac{\sum_{j=1}^{N} u_{ij}^2 (\mathbf{x_j} - \mathbf{v_i})(\mathbf{x_j} - \mathbf{v_i})^T}{\sum_{j=1}^{N} u_{ij}^2} \tag{2.14}$$

The Mahalanobis distance metric used by Gustafson and Kessel [66] and by Pedrycz and Waletzky [121] is a Fuzzy version because it takes into account the membership as well as the similarity between the data pattern and the cluster center. The inverse covariance matrix, $\mathbf{M_i}$ in (2.12) normalises dimensions of different scales, which prevents dominance from dimensions with greater scales. Thus, it is scale-invariant.

Kernel-based distance

The kernel methods solve non-linear problems by mapping the input space into higher dimensional space. Known as the 'kernel trick' and proposed by Aizerman *et al.* [5], they are applied to distances metrics by Schölkopf [137]. The idea here is to transform $\mathbf{x_j}$, a data point from a n-dimensional input space to a higher F-dimensional space resulting in $\Phi(\mathbf{x_j})$. The kernel-based distance between data pattern $\mathbf{x_j}$ and cluster centre $\mathbf{v_i}$ is defined as:

$$d_K^2(i,j) = ||\Phi(\mathbf{x_j}) - \Phi(\mathbf{v_i})||^2 \tag{2.15}$$

$$= K(\mathbf{x_j}, \mathbf{x_j}) - 2K(\mathbf{x_j}, \mathbf{v_i}) + K(\mathbf{v_i}, \mathbf{v_i}) \tag{2.16}$$

The Gaussian radial basis function can be used as a kernel function in the form:

$$K(a,b) = e^{\frac{-||a-b)||^2}{\sigma^2}} \tag{2.17}$$

where a and b are two data patterns and σ is a free parameter, or in this case, b is the cluster centre and σ is the standard deviation of all data patterns belonging to cluster i [165]. Thus, a kernel-based distance using Gaussian radial basis function yields a distance of the form:

$$d_K^2(i,j) = 2(1 - K(\mathbf{x_j}, \mathbf{v_i})) \tag{2.18}$$

2.3.6 Challenges of clustering

Jain [84] elaborated the problem of clustering due to the ambiguity in the definition of a cluster as well-separatedness does not always yield useful clusters. Furthermore, there is no definitive guide in the selection of data representation best suited for the defined clusters. The definition of a cluster and selection of the best data representation are determined by suitable choice of features, similarity measure, number of clusters and cluster validity [16, 81, 85], all of which cannot be easily determined.

According to Jain [84],"*... there is no universally good representation; the choice of representation must be guided by the domain knowledge*". Thus, the understanding of the data itself and what about the data to be retrieved is important. Data representation is determined by how similarities between data patterns are defined and by the choice of features with the domain knowledge in mind. This is further discussed by unifying the problem of clustering and the problem of identification of breast cancer subgroups through some of the existing work on finding useful breast cancer subgroups using clustering techniques in Chapter 2.8.

2.4 Semi-supervised Fuzzy c-Means

Semi-supervised learning is a hybrid of unsupervised and supervised learning, where both labelled and unlabelled data are used [30, 166]. There are many different types of semi-supervised learning settings; semi-supervised classification, constrained clustering, regression with labelled and unla-

belled data and dimensionality reduction [166]. The focus of this research is in semi-supervised classification and clustering using the ssFCM algorithm.

The first ssFCM algorithm was first introduced by Pedrycz [122]. He extended the objective function of the Gustafson and Kessel's FCM [66] to include supervised learning. Gustafson and Kessel's FCM uses the Fuzzy Mahalanobis distance metric previously discussed in Chapter 2.3.5. Pedrycz and Waletzky [121] improved the first ssFCM in [122] by modifying the scaling parameters α to ensure that the smaller population of labelled patterns can produce an impact in clustering. The resulting objective function contains unsupervised learning in the first term and supervised learning in the second term as follows:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^{p} d_{ij}^{2} + \alpha \sum_{i=1}^{c}\sum_{j=1}^{N} (u_{ij} - f_{ij} b_j)^{p} d_{ij}^{2}, \tag{2.19}$$

where

- u_{ij} is the membership value of data pattern j in cluster i,
- c is the number of clusters,
- d_{ij} the distance between data pattern j and cluster centre v_i,
- f_{ij} the membership value of labelled data pattern j in cluster i,
- b_j indicates if data pattern j is labelled,
- p is the fuzzifier parameter (which is commonly 2) and
- α is a scaling parameter for maintaining balance between the supervised and unsupervised learning components such that supervised learning does not dominate. The authors recommend α to be proportional to N/M, where M is the number of labelled data.

2.4.1 The Pedrycz and Waletzky [121] ssFCM algorithm

The objective function (2.19) is minimized using an optimization technique, i.e. the Lagrange multipliers, to derive equations for calculating the partition matrix. The prototypes are calculated as an average of the patterns with respect to the membership values. Like FCM, the algorithm iteratively calculates the cluster centres and the membership matrix $\mathbf{U}$ to minimise the objective function until a termination criterion is reached. Memberships of labelled and unlabelled data are updated and these memberships also contribute to the calculation of the cluster centres in (2.20). The algorithm is summarised as follows:

1. Initialise labelled data membership matrix $\mathbf{F}$ and initial membership matrix $\mathbf{U^0}$

2. Calculate the cluster centres $\mathbf{V} = [\mathbf{v_i}]$ with the partition matrix $\mathbf{U}$ using the following equation:

$$\mathbf{v_i} = \frac{\sum_{j=1}^{N} u_{ij}^2 \mathbf{x_j}}{\sum_{j=1}^{N} u_{ij}^2} \tag{2.20}$$

3. Update the partition matrix, $\mathbf{U}$ using the following equation:

$$u_{ij} = \frac{1}{1+\alpha} \left\{ \frac{1 + \alpha(1 - b_j \sum_{l=1}^{c} f_{lj})}{\sum_{l=1}^{c} (\frac{d_{ij}}{d_{lj}})^2} + \alpha f_{ij} b_j \right\} \tag{2.21}$$

4. If $||\mathbf{U}' - \mathbf{U}|| < \epsilon$, stop. Else, go to step 2 with $\mathbf{U} = \mathbf{U}'$

2.4.2 Developments and applications

Early ssFCM algorithms [122, 15, 121] have been shown to produce good classification results. Today, new ssFCM algorithms are developed to exploit FCM [21, 29, 155, 101, 47, 165, 94]. In this subsection, some of these developments and applications of ssFCM are reviewed.

Modification of scaling parameter α

The first ssFCM proposed by Pedrycz [122]. Pedrycz and Waletzky improved it by modifying the scaling parameter α [121]. The modification of objective function using α allows control over supervised and unsupervised learning of labelled and unlabelled data. Stutz and Runkler [146] have also modified the scaling parameter of the first ssFCM by using two separate parameters α and $(1-\alpha)$ to balance supervised and unsupervised learning respectively. Their modification is considered to be more general than the first ssFCM as labelled data undergo both supervised and unsupervised learning based on the two modified parameters α and $(1-\alpha)$.

Motivated by Stutz and Runkler's work, Li *et al.* [101] modified Pedrycz and Waletzkey's objective function [121] to reduce redundancy in unsupervised training. In [121], the unlabelled data patterns undergo unsupervised training twice. The modified objective function ensures unsupervised training is performed once for unlabelled data.

A low availability of labelled data would result in a higher α value (based on Pedrycz' and Waletzky's recommendation), giving a higher influence to the labelled patterns and therefore, there is more contribution from supervised learning. This is a good configuration given that the few labelled data are "good" to produce highly accurate classification. On the other hand, the presence of potentially poor labelled data will negatively affect clustering. Allowing some degree of unsupervised learning for labelled patterns may be a way to overcome this [121]. Bouchachia and Pedrycz in [23] observed that the scaling factor affected the well-separatedness of the clusters and concluded that assigning higher values to α indicates high confidence in the labelled data.

Introduction of weights

Bensaid *et al.* [15] applied ssFCM in image segmentation of Magnetic Resonance Images. They observed that the least square objective function of ssFCM tends to equalise cluster populations, which makes the representation of data unrealistic. Weights are introduced in the calculation of prototypes to counter this effect. The weights represent expert knowledge on the significance of each labelled patterns. Similarly, Pedrycz and Waletzky [121] introduce an additional weight termed as a *confidence factor* in the objective function instead of in the calculation of centroid [15]. A higher confidence factor for a pattern will result in a greater influence of the pattern. While this was proposed as a possible extension of (2.19), no supporting experiments have been presented.

The assignment of individual weights to labelled patterns is a tricky task as no definitive guide is found. In image segmentation, some visual guide can be obtained on how to assign weights based on areas of similar colours. In data clustering, the complexity of this task is increased with higher dimensions as it is not easily observable which labelled data is better.

Evolving membership

Bouchachia and Pedrycz [21, 23] modified the supervised component of Pedrycz and Waletzky's ssFCM algorithm [121] to counter the problem in ssFCM of patterns being inaccurately labelled by replacing the labelled pattern membership with evolving membership. The idea is to discriminate between accurately labelled patterns from those that are not. The evolving membership learns from the previous evolving membership and from the sum of differences between labelled pattern membership and previous evolving membership in different clusters for all patterns. The technique seems more computationally expensive because the evolving membership runs in a nested loop within each iteration.

Pairwise-constraints

Grira *et al.* [64] included supervision in the form of pairwise-constraints for pairs of patterns belonging to the same or different cluster/s in unsupervised FCM. The advantage of doing so is that experts can specify the constraints, which is then used to automatically calculate the membership, instead of manual or random initialisation. To tackle the number of clusters issue, Competitive Agglomeration introduced by Frigui and Krisnapuram [56] is incorporated, where a merging scheme based on cardinalities of the clusters is used. To further reduce the number of pairwise constraints used in their previous work [64], Grira *et al.* [65] incorporated active selection of constraints [12] into a ssFCM framework, allowing the selection to focus on the least well-defined clusters that are believed to give the most informative pairwise constraints. In these methods, however, it is unclear how the partition matrix is updated when clusters are merged.

Due to the nonlinear capabilities of kernel-based distance metric in ssFCM [165], Wang *et al.* [155] applied it in Grira *et al.*'s [64] pairwise-constrained competitive agglomeration (PCCA) ssFCM algorithm. According to them, the kernel-based pairwise-constrained technique could better utilize available constraints than PCCA and that using PCCA is impractical as a large range of parameter value has to be selected for different data sets. They further claimed that the imbalance between constraints and unlabelled patterns caused by using Euclidean distance increases the difficulty in this selection. Their approach was shown to outperform PCCA. However, discrepancies in PCCA's clustering results of common datasets in their paper [155] and the original paper [64] were found.

Motivated by the idea of preserving structure of neighbourhood in dimensionality reduction, Huang and Zhang [82] introduced a weighting system to preserve locality in the pairwise constrained ssFCM algorithm proposed by Grira *et al.* [65], but without competitive agglomeration. The

kernel-like weighting system is claimed to make the representation of the pattern more realistic by providing additional information about the structure of the neighbourhood. While comparison was made with other ssFCM for image segmentation, no comparison was made with other ssFCM algorithms for data clustering.

Distance metrics

In ssFCM algorithms [122, 121], the Fuzzy Mahalanobis distance was employed as it is able to represent non-linear data as compared to Euclidean distance metric and it is able to adapt to the shape of the clusters.

Zhang *et al.* [165] replaced the Euclidean distance metric in FCM with a kernel-based one because kernel functions could handle non-linear mapping functions without knowing the actual structure of these functions. The kernel-based distance metric is more desirable than the Euclidean distance metric for dealing with high dimensional data. In [23], Bouchachia and Pedrycz compared the effects of four different distance metrics on their proposed ssFCM algorithm with evolving membership introduced in [21]. They found that using kernelised distance produced the best results because new facts may be revealed and provide additional supervisory material to clustering. They stated that the true structure of the data may not be fully realised using Mahalonobis distance metric as there are only some labelled patterns to reflect this.

Höppner *et al.* [81] showed the different developments in distance metrics for modelling Fuzzy clusters to represent different shapes such as linear varieties or elliptotypes that best suit different datasets to solve application-specific problems. In [146], Stutz and Runkler applied Fuzzy c-Mixed prototypes in an ssFCM setting by replacing the distance metric with d_{prot_i} for the Fuzzy c-Mixed prototypes, which allows different types of prototypes to be represented in one dataset. This distance metric was developed as it

best represents the traffic data to classify and predict traffic statuses. For elliptotypes, d_ellip, the distance measure is defined as follows:

$$d_ik = \sqrt{||\mathbf{x_k} - \mathbf{v_i}||_A^2 - \alpha \sum_{j=1}^{r}((\mathbf{x_k} - \mathbf{v_i})^T A s_{ij})} \tag{2.22}$$

where

- $\alpha \in [0, 1]$ specifies locality, where $\alpha = 0$ indicates FCM (full locality) and $\alpha = 1$ for Fuzzy c-linear varieties (FCV) [18] (no locality).
- each s_{ij}, $i = 1, ..., c, j = 1, ..., r$ is a cluster director vector such that each cluster represents an r-dimensional linear subspace of $\mathbb{R}^p$. If $r = 1$, the algorithm can detect FCV Fuzzy c-lines (FCL) clusters.

Entropy regularisation

Endo *et al.* [47] introduced entropy regularised ssFCM objective function, which they claimed to be a simpler technique than Bouchachia and Pedrycz's ssFCM [21] technique since no evolving membership is required. The objective function of FCM is modified to include an entropy regularized term instead of a supervised FCM component. The study, however, did not provide comparative results with the other ssFCM techniques, nor provide experimental results tested on real datasets.

Summary

The developments in ssFCM that were discussed have a general aim of achieving a higher accuracy when compared to other existing ssFCM algorithms. There appears to be no clear, definitive guide as to when a particular approach should be used. There are application-specific proposals where modifications are made based on domain knowledge, such as those proposed for image segmentation and traffic flow prediction. However, there is no evidence that these proposed techniques perform favourably on

other datasets. One approach that performs favourably in some datasets may perform poorly in others, which suggests the need for experimental investigation in the exploration of a suitable approach for a dataset.

Furthermore, a standard evaluation methodology has not been established for ssFCM. Some [121, 101, 47, 64] have calculated the accuracy based on the number of matches from clustering solutions, while others use cross-validation [23], visual evaluation [15] or evaluation indices [141] such as Silhouette Index and Cohen's Kappa Index .

This research work is expected to be similar to one done by Tari *et al.* [148]. They applied ssFCM algorithm on biological datasets, two yeast microarray datasets to classify genes according to their gene functions using Gene Ontology annotations as prior knowledge. Another similar study is by Stutz and Runkler [146] where they demonstrated the use of ssFCM to identify meaningful subgroups in traffic data and to predict traffic statuses based on prior knowledge of these subgroups.

2.4.3 The motivation for study in semi-supervised Fuzzy c-Means

Jain [84] stated that one research direction of clustering is to achieve a tighter integration between clustering algorithms and application needs. This means that it is necessary to identify what the application requirements are and to tailor the clustering algorithm based on these requirements. One fundamental issue of clustering which he raised was the consistency of solutions from different clustering algorithms, that is, the stability of these different clustering solutions. He explained the use of semi-supervised clustering is beneficial in deciding (i) data representation and (ii) appropriate objective function for data clustering using labelled data and (user-specified) pair-wise constraints.

Based on Mackay's [105] and Jain's [84] discussions on clustering and the nature of the NTBC dataset, the focus in the ssFCM clustering al-

gorithm is chosen. Being semi-supervised, ssFCM can use both labelled and unlabelled data during training. In addition, being FCM-based, ssFCM allows data patterns to be represented in more than one clusters using Fuzzy memberships. Although ssFCM is a clustering technique, it has been demonstrated to perform classification tasks successfully using class labels [121, 101, 146] with the number of clusters equals the number of classes. In this work, the practice of setting the number of clusters c to the six classes identified by Soria *et al.* [141] is adopted. ssFCM has also been successfully applied in areas of biomedicine [29, 148], where boundaries of classes are often unclear. The ability of ssFCM to represent data in more than one cluster using membership values makes it highly suitable for classifying biomedical dataset where the unclear boundary of classes can be represented using Fuzzy membership. The ssFCM algorithm which is intended to be applied can learn from labelled data provided from Soria's classification [141]. Another benefit of using ssFCM is that clustering is not a statistical inference technique and therefore, is not affected by the assumptions of normal distribution [69], which is suitable on the NTBC dataset whose features are not normally-distributed [140]. ssFCM algorithms have been demonstrated to perform classification tasks in both a clustering and a classification setting, often with the number of clusters equals the number of classes [121, 101, 23]. In a clustering setting, the ssFCM algorithm is run once while in a classification setting, the algorithm is first trained and then tested [23].

2.5 Evaluation techniques

As a wide range of evaluation techniques are used across different literatures in ssFCM, some is achieved in the following sections. The techniques reviewed are used for experiments and were chosen based on their application in evaluating clustering and/or classification techniques [121, 141, 23, 109].

2.5.1 Accuracy rate

Accuracy rate is the average number of matches between the cluster labels and class labels of data patterns over total number of comparisons. It ranges between values 0 and 1 where 0 indicates no agreement and 1 indicates complete agreement.

2.5.2 Cohen's Kappa index

The Cohen's Kappa index (κ) [35] is given by:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the ratio of agreements between the two sources and p_e is the ratio of chances of agreement. The p_e ratio is calculated based on the sum of the probabilities of the sources having agreements randomly. A κ value of 0 indicates no agreement while 1 indicates complete agreement.

2.5.3 Normalised Mutual Information

Normalised Mutual Information (NMI) [145] calculates the comparison of clustering solutions in terms of cluster-class (or cluster-cluster) matching and distribution and normalises this calculation. The NMI equation is defined as follows:

$$\text{NMI}(X, Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \tag{2.23}$$

where $I(X;Y)$ denotes Mutual Information between variables X and Y and $H(X)$ and $H(Y)$ denote the entropy of variables X and Y respectively. $I(X;Y)$ is computed as follows:

$$I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X) \tag{2.24}$$

where $H(X|Y)$ and $H(Y|X)$ are conditional entropies and H(X,Y) are joint entropy. NMI values close to zero denotes low agreement while a near 1 value indicates otherwise.

2.5.4 Cross-validation

Cross-validation is used to evaluate the accuracy of a classification technique. It first divides the data into training and test sets. The training set is used as examples to learn the best model. The model obtained is then used to predict the labels of the data pattern in the test sets.

One common type of cross-validation is the k-fold cross validation (CV) where the dataset is divided into k equally-sized subsets such that k-1 sets are used as the training set and the remaining as the test set. The process is repeated k times, with each of the k sets being used as test set at different runs of the training process. In this way, all data patterns are used for training and testing without being repeatedly used in both and they are validated exactly once.

The accuracy rate tends to give a more optimistic view than the Kappa Index and NMI because it only takes into account of the agreements and completely disregards the disagreements. Both Kappa Index and NMI take into account of the agreements and disagreements where there is some sort of penalty for disagreements. In NMI, $H(X|Y)$ and $H(Y|X)$ represent the disagreements. In Kappa, the disagreements are taken into account in the form of a probability of random agreement p_e. Cross-validation allows the evaluation of classification tested on unseen data, which is more realistic as the model built is being tested, and can be incorporated with the other evaluation technique. Furthermore, it reduces overfitting.

2.6 Initialisation techniques

In clustering algorithms such as K-means and FCM, the common practice of random initialisations often lead to different suboptimal solutions in different runs [83]. In this section, some of the existing initialisation techniques that can refine clustering methods are reviewed. These techniques have been shown to improve clustering techniques such as K-means or FCM [78, 32], but, to the best of our knowledge, no investigative studies on using these techniques to improve ssFCM has been done.

2.6.1 Cluster Estimation

Proposed by Chiu [32], the Cluster Estimation (CE) technique estimates both the number and location of cluster centres by specifying its neighbourhood size r. Based on the number of neighbouring patterns, a potential value is calculated as follows:

$$P_i = \sum_{j=1}^{n} e^{-\alpha||\mathbf{x_i}-\mathbf{x_j}||^2} \tag{2.25}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are two data patterns and $\alpha = \frac{4}{r_a^2}$. The pattern with the highest potential value becomes the first cluster centre. Eq. (2.25) is then revised to calculate the potential of patterns to be centres of other clusters, as shown below:

$$P_i \Leftarrow P_i - P_k^* e^{-\beta||\mathbf{x_i}-\mathbf{x_k^*}||^2} \tag{2.26}$$

where $\beta = \frac{4}{r_b^2}$, $\mathbf{x_k^*}$ is the latest obtained cluster centre and P_k^* its potential calculated from (2.25). The positive constants r_a and r_b are radius defining their respective neighbourhoods. The author recommended that r_b=1.25r_a. In CE, the selection of the influence parameter becomes tricky in datasets with highly overlapping clusters. Selecting a large range of influence value r, causes similar neighbourhoods for each data patterns to form, causing less variation in potential value of data patterns. If the value is too small,

on the other hand, neighbours which could contribute to the potential value of the data pattern (for it to qualify as cluster centre) may be disregarded. CE is able to disregard noisy data by limiting its radius of influence. Noisy data within this radius will fetch poor potential values and therefore, will never be selected as centres.

2.6.2 Simple Cluster Seeking

The Simple Cluster Seeking (SCS) technique defined by Tou and Gonzales [153] is summarised as follows (as described by He *et al.* [78]:

1. The first pattern is initialised as the first cluster centre, i.e. $\mathbf{v_1} = \mathbf{x_1}$.

2. For $k = 2, ..., N$, $\mathbf{x_k}$ is the next cluster centre if $||\mathbf{x_k} - \mathbf{v_i}|| > \rho$ for all existing cluster centres, where ρ is a parameter specifying the distance between two cluster centres. When c cluster centres are initialised, stop. Else, decrease the value of ρ and repeat the steps.

SCS is sensitive to initial ρ value and the order of the data patterns. It picks the first data pattern whose distance to all the cluster centres is larger than a threshold value and this process repeated until all centres are found. This means the next data pattern may be a better candidate as a cluster centre, but would not be chosen.

2.6.3 Katsavounidis *et al.* initialisation

The initialisation technique by Katsavounidis, Kuo and Zhang (KKZ) [90] takes on the following steps as described in [6]:

1. Initialise the first cluster centre with the data pattern that has the maximum norm, $\mathbf{v_1} = \arg\max ||\mathbf{x_k}||$.

2. Initialise the second cluster centre with the data pattern furthest from $\mathbf{v_1}$.

3. Compute the minimum distances between the remaining points with all initialised cluster centres. The data pattern with the largest value of these minimum distances are chosen as the next cluster centre.

4. Repeat step 3 until all cluster centres are found.

KKZ tends to choose centres located at the edge of the cluster. In the presence of noisy data, KKZ's choice in the centres becomes affected. This means KKZ is highly susceptible to noise as it regards the noisy data as part of the cluster. Despite its drawbacks, it is simple and fast [78].

Clustering techniques such as FCM or K-means can be used to initialise clustering algorithms where the cluster centres or clustering solutions are used [118, 101, 13].

2.7 Feature selection

While it makes sense that having more features give more discriminating power to distinguish between classes, in practice, more features increase time requirements. Furthermore, the irrelevant or redundant features can worsen classification accuracy. Hall [71] has described these features as "harmful redundancies". So far, despite the high popularity in feature selection, only a few studies have applied them in combination with ssFCM. Benkhalifa and Bensaid applied a filter-based feature selection technique using Information Gain with ssFCM in categorising text [14]. Park and Yae [120] used ssFCM and Support Vector Machines to select and evaluate features.

Feature selection is of interest to this research work for two reasons. First, it has been shown to improve classification of learning algorithms [71, 68]. Secondly, in the case of the NTBC dataset, by reducing the number of protein biomarkers while maintaining classification accuracy, the time and cost to run clinical tests for data collection are reduced. Furthermore, the

most relevant (important) protein biomarkers can be identified and in turn, subgroup assignment of future breast cancer patients can be determined using this reduced panel of the most relevant protein biomarkers.

Many feature selection techniques have been proposed. A general review of the topic has been covered by Dash and Liu [39] and Saeys *et al.* [135]. Furthermore, Hall's thesis [71] provides a thorough introduction to feature selection. The following feature selection techniques which are used in this research work are reviewed; Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [68], Random Forest-RFE (RF-RFE) [62, 95] , Naive bayes-RFE (NB-RFE) [95] and Correlation-based Feature Selection (CFS) by Hall [72, 71]. These techniques are further discussed next.

2.7.1 Selected techniques

There are three types of feature selection methods; filter, wrapper and embedded [67]. The filter method uses heuristics to evaluate the feature subsets. Wrappers use machine learning algorithms based on CV to evaluate the feature subsets. They repeatedly search through the feature subsets find the subset that best fit the classification model. The search can be based on best first search or greedy search. Embedded methods use machine learning and ranking technique based on weights (generated by the machine learning algorithm) to select features during training. These methods are illustrated in Figure 2.1, as in [71, 135]. The CFS algorithm is a filter method while NB-RFE and RF-RFE are wrappers and SVM-RFE is an embedded method. These techniques are reviewed as follows.

Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) [72] searches through the space of feature subsets and evaluate the goodness of the feature subsets using a heuristic. The heuristic takes into account the usefulness of each features

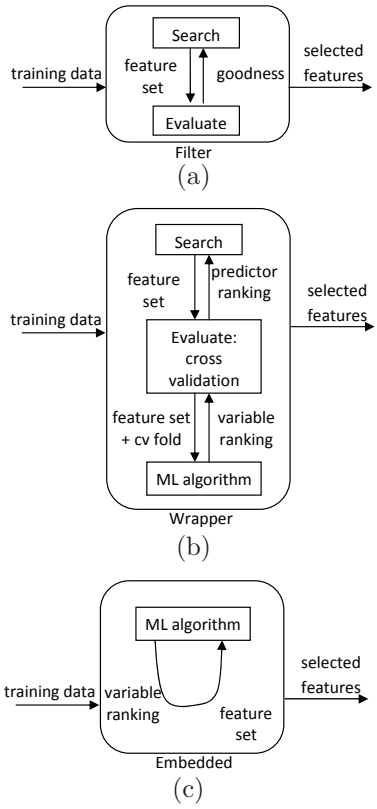


Figure 2.1: The three types of feature selection: filter, wrapper and embedded

based on how well it predicts class labels and the intercorrelation among them. It is formulated as follows:

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \tag{2.27}$$

where r_{zc} is the correlation between summed components and the outside variable. The components are data from a set of selected features that goes in a test to measure traits related to an outside variable (class). k is the number of components (features), $\overline{r_{zi}}$ is the average of the correlations between the components and the outside variable and $\overline{r_{ii}}$ is the average inter-correlation between components. The idea is to choose feature subsets with maximum r_{zc}.

Wrapper-based Recursive Feature Elimination

The recursive feature elimination (RFE) is an iterative backward selection method which removes the feature based on a ranking criteria. It can be

Algorithm 1 Wrapper-based RFE using resampling [95]

```
for Each Resampling Iteration do
    Partition data into training and test sets via resampling
    Train model on training set using all predictors with a machine learning algorithm
    Predict using the test set
    Calculate feature ranking
    for Each subset size S_i, i = 1...S do
        Keep S_i most important features
        Train model on the training set using S_i predictors
        Predict the test set
        Calculate ranking of each predictor
    end for
end for
Calculate performance over S_i using test set
Determine the appropriate number of predictors
Determine list of predictors to keep in the final model
Fit the final model based on optimal S_i using original training set
```

used in a wrapper-based or in an embedded method. In a wrapper-based method, the algorithm is shown in Algorithm 1 as described in [95] (for the `rfe` function in R).

S is a sequence of ordered candidate values for the number of predictors to be retained such that $S_1 > S_2, ...$ The S_i top ranked predictors are retained and the model is refitted and evaluated. In this way, features are assessed on whether they are useful to be predictors. Wrapper methods are prone to overfitting as the algorithm may focus on characteristics of training data that are not found in future data. Thus, resampling (using CV, for instance) is introduced in Algorithm 1 to take into account of variability in the data. Despite being computationally expensive, RFE is expected to produce higher accuracy. Classifiers such as Naive Bayes (NB-RFE) and Random Forests (RF-RFE) are used to train the models.

Support Vector Machine - Recursive Feature Elimination

The Support Vector Machine - Recursive Feature Elimination (SVM-RFE) technique was proposed by Guyon *et al.* [68] and is based on RFE using

Algorithm 2 SVM-RFE [68]

1: **for** Subset of surviving features $\mathbf{s} = [1, 2, ...n]$ **do**
2: Restrict training examples $\mathbf{X_0} = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_k, ...\mathbf{x}_\ell]^T$ to good feature indices: $\mathbf{X} = \mathbf{X_0}(:, \mathbf{s})$
3: Train the classifier using SVM where $\mathbf{y} = [\mathrm{y}_1, \mathrm{y}_2, ...\mathrm{y}_\mathrm{k}, ...\mathrm{y}_\ell]^\mathrm{T}$ are the class labels: $\alpha = \mathrm{train}(\mathbf{X}, \mathbf{y})$
4: Compute the weight vector of dimension length ($\mathbf{s}$): $\mathbf{w} = \sum_\mathrm{k} \alpha_\mathrm{k} \mathrm{y}_\mathrm{k} \mathbf{x}_\mathrm{k}$
5: Calculate the ranking criteria: $\mathrm{c_i} = (\mathrm{w_i})^2$, for all i
6: Find the feature with the smallest ranking criterion: $\mathrm{f} = argmin(\mathbf{c})$
7: Update feature ranked list: $\mathbf{r} = [\mathbf{s}(\mathrm{f}), \mathbf{r}]$
8: Eliminate the feature with smallest ranking criterion: $\mathbf{s} = \mathbf{s}(1 : \mathrm{f} - 1, \mathrm{f} + 1 : \mathrm{length}(\mathbf{s}))$
9: **end for**

the weight magnitude of SVM as a ranking criterion. The algorithm is described in Algorithm 2, where the output is a feature ranked list $\mathbf{r}$.

2.7.2 Issues

There are two issues in feature selection; overfitting and stability of selected features. Ng [115] warns of the dangers of overfitting from using cross-validation data. According to Guyon and Elisseeff [67], wrapper-based techniques are prone to overfitting. The cross-validation procedure used in wrapper-based techniques may have attributed this, such that the selected features will produce improved classification only for some classification techniques. To prevent overfitting in wrapper-based techniques, Kuhn [95] introduced resampling in the outer loop in Algorithm 1 so that different training and test sets are used in each resampling.

According to Kalousis *et al.* [89], to select a suitable feature selection technique to use with a classification technique, one approach is to use the combination of feature selection and classifier with the most stable feature selection technique. Stability measures have been used as evaluation of feature selection. If a feature selection technique consistently selects the same features, it builds confidence in the importance of selected features.

With higher number of selected features, higher stability is expected as there is higher probability of selecting common features. The stability measure by Kalousis *et al.* [89] measures the amount of overlapping features between two subsets of features. It is based on the Tanimoto distance [43] and defined as:

$$S(s, s') = 1 - \frac{|s| + |s'| - 2|s \cap s'|}{|s| + |s'| - |s \cap s'|} \tag{2.28}$$

where s and s' are subsets of features. S values are in [0,1] where 0 indicates no overlap and 1 indicates the two subsets are the same.

2.8 Clustering for breast cancer classification

Prognosis is the prediction of the likely outcome of one's current medical condition such as survival rate or survival time. Current prognosis techniques involve using clinical and pathologic prognostic and predictive factors [136, 141]. According to Cianfrocca and Goldstein [34], prognostic factors are measurements obtained during surgery that relates to the disease-free or overall survival in the absence of therapy. These factors, thus, can relate with the natural history of the disease. Predictive factors, on the other hand, are measurements associated with response to a given therapy. There are also factors that are both prognostic and predictive, such as hormone receptors, HER2/neu overexpression and gene expression profiles. Bundred [28] further explains that prognostic factors are either chronological, which indicates how long the cancer has been present, such as tumour size; or biological, which indicates the metastatic potential behaviour of a tumour, such as tumour grade. To determine prognosis, a prognosis index for primary breast cancer was constructed based on prognosis factors, lymph-node stage, tumour size and pathological (tumour) grade [76, 57]. McGuire [111] identified a set of criteria to ensure that a

prognostic factor have clinical relevance. It, therefore, must have biological relevance, be reproducible, be validated using many patients, be independently confirmed by other experts and have optimised cutoff values.

Clustering of biomedical data is becoming popular and increasingly important in assisting clinicians in the identification of specific types of illnesses or diseases. These techniques, therefore, can support decision making in prognosis and treatment [154, 1]. Eisen *et al.* [46] showed that gene expression data can be organised into functional categories using hierarchical clustering and visual inspection of the dendrogram. The identification of functional categories from clustering gene expression data motivated the application of clustering algorithms on breast cancer gene expression data [124] where Perou *et al.* identified four breast cancer subgroups: ER+/luminal-like, basal-like, HER2+ and normal breast. In a following study [143], Sorlie *et al.* identified six subgroups where the ER+/luminal-like group was further divided into three subgroups: luminal-A, B, and C. Luminal-C was dropped in a later work [144] but, the reason for it is unclear. Researchers moved on to cluster immunohistochemical data using hierarchical clustering, where three subgroups were identified in [108] and six subgroups were determined in [3]. However, there has been no further investigations to address the stability of the proposed subgroups. The reproducibility of these subgroups, therefore, has not been assessed using different breast cancer datasets, such as those in [144], or using different learning algorithms.

Clustering biomedical data is a complex procedure as the subgroups found may not be correct or meaningful. Bair and Tibshirani [9] explained the difficulty and importance of finding relevance between subgroups and clinical parameters for accurate prognosis. As unsupervised approaches often do not use clinical data to find subgroups, there is no assurance that the subgroups found would be related to the clinical outcome. Furthermore, in

some cases the subgroups identified from clinical data may not be biologically meaningful. Statistical tests are, therefore, needed to determine the relevance between the classification and clinical data, which in turn can validate the subgroups identified.

Therefore, in order to validate the biomedical subgroups, the subgroups have to be biologically meaningful and reproducible. This means that relevance between the subgroups and clinical parameters have to exist and the stability of these subgroups has to be addressed.

2.8.1 The Nottingham Tenovus Breast Cancer dataset

The Nottingham Tenovus Breast Cancer (NTBC) dataset contains immunohistochemical data of 1076 patients with primary operable (stages I, II and III) invasive breast cancer between 1986 and 1998. The data is in the form of modified histochemical score (H-score) based on immunohistochemical reactivity of 25 proteins, determined using microscopical analysis. The H-score is calculated based on a semiquantitative assessment of both intensity of staining and percentage of positive cells at each intensity. The intensity of staining is scored 0 to 3, which correspond to negative, weak, moderate and strong positivity. The H-score ranges between 0 and 300, based on the formula:

$$\begin{aligned} \text{H-score} &= (1 \times \% \text{ of cells with intensity 1}) \\ &+ (2 \times \% \text{ of cells with intensity 2}) \\ &+ (3 \times \% \text{ of cells with intensity 3}) \end{aligned} \tag{2.29}$$

The 25 protein biomarkers (features) are the same ones listed in [141, 3] and are shown in Table 2.2 on page 49. The dataset also contains clinical data such as histologic grade, histologic tumour type, vascular invasion, tumour size, lymph node stage, patient age and menopausal status. Survival (in months) from the date of primary treatment to the time of death

is recorded at 3-months intervals initially, then every 6 months, and finally, annually for a range of 1-192 months, with a median period of 58 months. The Nottingham Prognostic Index (NPI) [57] score is also recorded. It is calculated based on prognostic factors according to the formula:

$$\text{NPI Score} = (0.2 \times \text{tumour size}) + \text{histologic grade} + \text{lymph node stage} \tag{2.30}$$

A poor prognosis is indicated by a high NPI score. Table 2.3 on page 49 shows the NPI ranges and their interpretations.

2.8.2 Discovery of subgroups in the dataset

The Nottingham Tenovus Breast Cancer (NTBC) dataset has been clustered using hierarchical clustering into five subgroups, with the sixth subgroup containing only four patients [3]. To address the stability of the proposed groups in NTBC, Soria *et al.* [141] first identify the optimal number of clusters in the dataset using different cluster validity indices on clustering solutions from K-means and Partitioning Around Medoids (PAM). They found that the most stable number of clusters obtained by PAM is four, and six by K-means. Then, a set of rules is used to determine consensual classes with solutions obtained from hierarchical clustering, k-means and Adaptive Resonance Theory (ART). The key biomarkers that characterise these classes are identified [141] using visual inspection, Orthogonal Search Rule Extraction (OSRE) [49] and Artificial Neural Networks (ANN) [113]. To automate the process of identifying the same six classes and thereby address the stability issue, Soria *et al.* [140] used three different classifiers, C4.5 [126], Multi Layer Perceptron- Artificial Neural Networks (MLP-ANN) [77] and Naive Bayes [88].

Based on classification by Soria *et al.*, there are three main clinical groups, Luminal, Basal and HER2. These main groups are further divided into six subgroups where class 1, 2 and 3 belong to the Luminal group.

Table 2.2: Protein biomarkers and their dilutions.

Antibody, clone	Short name	Dilution
Luminal phenotype		
CK 7/8 [clone CAM 5.2]	CK7/8	1:2
CK 18 [clone DC 10]	CK18	1:50
CK 19 [clone BCK 108]	CK19	1:100
Basal phenotype		
CK 5/6 [clone D5/16134]	CK5/6	1:100
CK 14 [clone LL002]	CK14	1:100
SMA [clone 1A4]	Actin	1:2000
p63 ab-1 [clone4A4]	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
c-erbB-2	HER2	1:250
c-erbB-3 [clone RTJ1]	HER3	1:20
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 2.3: Interpretation of the Nottingham Prognosis Index

NPI score	Interpretation
$\leqslant 2.4$	Excellent Prognosis(EPG)
$2.4<NPI\leqslant 3.4$	Good Prognosis (GPG)
$3.4<NPI\leqslant 4.4$	Moderate Prognosis 1 (MPG1)
$4.4<NPI\leqslant 5.4$	Moderate Prognosis 2 (MPG2)
<5.4	Poor Prognosis (PPG)

Table 2.4: Number of data patterns in each class and the number of not classified (n.c) and classified (c) data patterns according to classification by Soria *et al.* [141]

class 1	class 2	class 3	class 4	class 5	class 6	n.c	c
202	153	80	82	69	77	413	663

Class 4 and 5 belong to the Basal group and class 6 to HER2. Each class is named (in square brackets) and described by key features identified by Soria *et al.* [141] as follows:

- class 1 [Luminal A]: ER+, PgR+, CK7/8+, CK18+, CK19+, HER3+, HER4+
- class 2 [Luminal N]: ER+, PgR+, CK7/8+, CK18+, CK19+, HER3-, HER4-
- class 3 [Luminal B]: ER+, PgR-, CK7/8+, CK18+, CK19+, HER3+, HER4+
- class 4 [Basal - p53 altered]: ER-, p53+, CK5/6+, CK14+
- class 5 [Basal - p53 normal]: ER-, p53-, CK5/6+, CK14+
- class 6 [HER2]: ER-, HER2+

As shown on Table 2.2 on page 49, ER and PgR are hormone receptors. CK7/8, CK18 and CK19 are luminal cytokeratins. CK5/6 and CK14 are basal cytokeratins. HER2, HER3 and HER4 are EGFR family members. p53 is a tumour suppressor gene. The + or - at the end of each feature indicates high or low expressions respectively. In Soria's classification [141], 663 data patterns are classified while 413 remains not classified (n.c) , as shown in Table 2.4. While the consensus clustering methodology proposed by Soria *et al.* has identified six clinically useful subgroups, this is at the expense of not classifying the entire dataset where the 413 data patterns (patients) belonging to mixed classes are not assigned to any class.

In [142], Soria *et al.* proposed a quantifier-based classification system with a reduced panel of biomarkers to refine the previous classification in [141], classifying the entire dataset into seven subgroups. Based on the refined classification and same reduced panel of biomarkers, clinical association between the seven subgroups were further examined by Green *et al.* [63]

and key clinical breast cancer phenotypes for these subgroups were identified. By stratification of these seven biological subgroups using NPI-like formulae on clinical parameters (called NPI+), Rakha *et al.* showed that distinct prognostic groups can be identified [132]. Thus, the NPI has evolved from using only clinical data to NPI+ which uses both biological and clinical data for prognosis.

2.9 Summary

The development of ssFCM methodologies for application on real world biomedical datasets is important. Semi-supervised clustering techniques such as ssFCM are ideal for datasets where the availability of labels are scarce and/or the collection of labels is complex and expensive. Furthermore, ssFCM algorithms are capable of representing data patterns in biomedical datasets where boundaries between subgroups are not clearly defined with the use of memberships to indicate belongingness to more than one cluster. They can also learn from available labelled data to identify similar data patterns. In addition, the amount of labelled data can be adjusted, which allows control over the amount of influence from labelled data in ssFCM. If similar subgroups (previously found) can be reproduced using ssFCM with some labelled data, the stability of the subgroups can be demonstrated, which is a form of validation for the subgroups previously identified. Furthermore, ssFCM have been shown to produce good classification results on popular UCI and real-world datasets [121, 101, 15, 23].

The ssFCM algorithm can be a single-clustering-algorithm approach to classify the dataset with guidance from Soria's classification [141]. The identification of similar biological subgroups as those found by Soria *et al.* can further validate these subgroups. It has been established that there should be relevance between subgroups found from clustering algorithms in biomedical data and clinical parameters in order for the subgroups be

be considered clinically useful. Furthermore, the subgroups themselves should be biologically meaningful and this can be validated with existing findings in related literatures. No one clustering technique provides the best solution for all datasets. It is, therefore, vital to first understand the dataset and the issues of considered techniques and subsequently investigate their suitability through experimentation.

Furthermore, initialisation techniques can be applied to ssFCM to provide a better start for the algorithm than solely relying on Soria's classification [141]. The hypothesis of this research is that the ssFCM can be successfully applied on the NTBC dataset to identify stable breast cancer subgroups, achieving good agreement (with percentage matches of 80%) with Soria's classification. In addition, the agreement with Soria's classification can be further increased by incorporating other approaches such as initialisation techniques and feature selection into ssFCM. Therefore, in the next chapters, the investigations in ssFCM and in the application of other approaches to improve ssFCM clustering or classification on the NTBC dataset are described, given the research gap found.

3 Preliminary Studies

In this chapter, three investigations aimed at determining whether ssFCM is suitable for clustering (identification of subgroups) or for classifying (prediction of class labels for new patients) the NTBC dataset are conducted. The first investigation is an exploratory study that compares clustering performance of different ssFCM algorithms applied to popular UCI datasets [54]. As there are a number of different types of ssFCM, the distance-based ones are focused on because the first ssFCM by Pedrycz [122] is of this type and many existing ssFCM algorithms [121, 165, 101, 47] are extended forms of this type. This study will help to uncover the factors that affect accuracy of ssFCM and to identify the most suitable ssFCM for application on the NTBC dataset. Based on experimental results, the ssFCM algorithm by Pedrycz and Waletzky [121] is the most favourable as they produced the highest accuracy in a majority of UCI datasets. Issues such as scale differences of dimensions, distance metrics, objective functions and quality of labelled patterns have been observed to affect classification results.

The second investigation explores different distance metrics with the selected ssFCM from the previous investigation, the algorithm by Pedrycz and Waletzky [121], to determine the most suitable distance metric for representing the NTBC dataset. The ssFCM by Pedrycz and Waletzky with Euclidean distance (which shall be referred as ssFCM in the rest of the thesis for short) is observed to produce the highest average accuracy than with Mahalanobis, Fuzzy Mahalanobis and kernel-based distance.

The third investigation is a comparative study to determine how well ssFCM classifies the NTBC dataset in comparison with other classifiers such as k Nearest Neighbours and Naive Bayes. The experimental results,

based on a 10-fold CV, showed that ssFCM produced results with the highest similarity to Soria's classification.

The experimental results from these studies demonstrated that ssFCM can successfully classify popular datasets and with selection of a suitable distance metric, ssFCM can classify the NTBC dataset with high agreement to Soria's classification [141], outperforming other popular classifiers.

3.1 A comparative investigation in distance-based semi-supervised Fuzzy c-Means

3.1.1 Background and motivation

Different ssFCM algorithms have previously been evaluated on different datasets using various ranges of labelled patterns, which makes fair comparison difficult. This motivated the study into a comparison of four distance-based semi-supervised Fuzzy c-Means (FCM) algorithms proposed by Pedrycz and Waletzky [121], Zhang *et al.* [165], Li *et al.* [101] and Endo *et al.* [47]. These were chosen as they were found to be simple in operation and have been shown to produce good results. The objectives of this section are (i) to compare their performances with varying quantities of labelled patterns, and (ii) to explore how common issues to semi-supervised Fuzzy clustering affect their performance. These issues include the number of dimensions, application on different datasets, choice of initial membership values, objective functions, distance metrics and labelled patterns. Some of these issues have been discussed in [23] where different distance metrics of similar objective functions were investigated. However, in this section, comparisons are made between algorithms with different objective functions with different distance metrics, and algorithms with similar objective functions that employed different forms of balance between supervised and unsupervised learning, as described next.

Table 3.1: Pedrycz-97 algorithm [121]

Objective Function	$J_k = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{p} d_{ik}^{2}$ $+\alpha \sum_{i=1}^{c}\sum_{k=1}^{N} (u_{ik} - f_{ik}b_k)^{p} d_{ik}^{2}$
Centroid	$\mathbf{v_i} = \frac{\sum_{k=1}^{N} u_{ik}^{2}\mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^{2}}$
Partition Matrix	$u_{ij} = \frac{1}{1+\alpha}\left\{ \frac{1+\alpha(1-b_j\sum_{l=1}^{c} f_{lj})}{\sum_{l=1}^{c}(\frac{d_{ij}}{d_{lj}})^2} + \alpha f_{ij} b_j \right\}$
Distance Metric	Fuzzy Mahalonobis

Table 3.2: Li-08 algorithm [101]

Objective Function	$J_m = \sum_{i=1}^{c}\sum_{k=L+1}^{N} u_{ik}^{m} d_{ik}^{2} + (1-a)\sum_{i=1}^{c}\sum_{k=1}^{L} u_{ik}^{m} d_{ik}^{2}$ $+a\sum_{i=1}^{c}\sum_{k=1}^{L} (u_{ik} - f_{ik})^{m} d_{ik}^{2}$
Centroid	$\mathbf{v_i^{(l)}} = \frac{\sum_{k=1}^{N} (u_{ik}^{(l)})^2 \mathbf{x}_k}{\sum_{k=1}^{N} (u_{ik}^{(l)})^2}$
Partition Matrix	If unlabelled, $u_{ij} = \frac{1}{\sum_{l=1}^{c}(\frac{d_{ij}}{d_{lj}})^2} = u_{ij}^{FCM}$
	If labelled, $u_{ik} = (1-a)\left(\frac{1}{\sum_{l=1}^{c}(\frac{d_{ij}}{d_{lj}})^2}\right) + a f_{ik}$ $= (1-a)u_{ik}^{FCM} + a f_{ik}$
Distance Metric	Fuzzy Mahalanobis

Table 3.3: Zhang-04 algorithm [165]

Objective Function	$J_m = 2\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m}(1 - K(x_k, v_i))$
Centroid	$\mathbf{v_i} = \frac{\sum_{k=1}^{N_l}(u_{ik}^{l})^m K(x_k^l, v_i) x_k^l + \sum_{k=1}^{N_u}(u_{ik}^{u})^m K(x_k^u, v_i) x_k^u}{\sum_{k=1}^{N_l}(u_{ik}^{l})^m K(x_k^l, v_i) + \sum_{k=1}^{N_u}(u_{ik}^{u})^m K(x_k^u, v_i)}$
Partition Matrix	$u_{ik}^{u} = \frac{(1/(1-K(x_k^u, v_i)))^{1/(m-1)}}{\sum_{j=1}^{c}(1/(1-K(x_k^u, v_j)))^{1/(m-1)}}$, $1 \leqslant i \leqslant c, 1 \leqslant k \leqslant N_u$
Distance Metric	Kernel-based distance

3.1.2 The selected algorithms

Tables 3.1, 3.2, 3.3 and 3.4 illustrate the components in the ssFCM algorithms selected for study; Pedrycz-97 [121], Li-08 [101], Zhang-04 [165] and Endo-09 [47], respectively. The equations for calculating the centroid and partition matrix are found in these tables. They are derived using a standard optimization technique called the Lagrange multipliers, the derivations are found in their respective papers and will not be discussed here as they are not within the scope of the research aims.

Table 3.4: Endo-09 algorithm [47]

Objective Function	$J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik} \|\|x_k - v_i\|\|^2$ $+\lambda^{-1} \sum_{i+1}^{c} \sum_{k=1}^{N} (u_{ik} - \bar{u}_{ik}) log(u_{ik} - \bar{u}_{ik})$
Centroid	$v_i = \frac{\sum_{k=1}^{N} u_{ik} x_k}{\sum_{k=1}^{N} u_{ik}}$
Partition Matrix	$u_{ik} = \bar{u}_{ik} + \frac{e^{-\lambda d_{ik}}}{\sum_{j=1}^{c} e^{-\lambda d_{jk}}} \left(1 - \sum_{j=1}^{c} \bar{u}_{jk}\right)$
Distance Metric	Euclidean

Differences

Both Pedrycz-97 and Li-08 use Fuzzy Mahalonobis distance to update their partition matrix. In the objective function of Pedrycz-97, all labelled data participate in unsupervised and supervised learning with parameter α to maintain a balance between the two types of learning. α is recommended to be proportional to N/M where M is the number of labelled patterns. Li-08 is an improvement to avoid the redundant unsupervised learning for labelled patterns in Pedrycz-97. Instead, it takes all unlabelled patterns and $(1 - a)$ of labelled patterns to participate in unsupervised learning, and a of labelled patterns to undergo supervised learning. Li *et al.* recommended that $a = 1 - M/N$. Though presented separately for labelled and unlabelled patterns in Li-08, the equation to update the partition matrix is similar to Pedrycz-97 except for their balance parameters, a and α, respectively. In Li-08, the unsupervised learning of labelled data are reduced, thus more reliance on labelled data. Unlabelled patterns are not used in the calculation of Li-09's centroids, unlike the other algorithms. Like Pedrycz-97, the objective function of Endo-09 trains both labelled and unlabelled patterns in both unsupervised and supervised fashion, but using the Euclidean distance metric. The supervised training function, however, is entropy-regularised. Zhang *et al.* retained the objective function from the original FCM, but replaced the Euclidean distance metric with a Gaussian kernel-based one. Unlike Pedrycz-97 and Li-08, in Zhang-04 only unlabelled patterns undergo supervised learning. This means that

labelled patterns never gets updated or improved. As the unlabelled patterns do not contribute to the initial centroid, the labelled patterns act as seeds to form initial clusters, applied in ssFCM learning [102] and non-Fuzzy semi-supervised learning [11]. The trouble with using membership values of labelled patterns in both the calculation of centroids and the partition matrix, is that it makes the algorithm highly dependent on the initial membership values of labelled patterns, which can negatively impact the classification results if they contain errors. This highlights the importance of balance (scaling) parameters like α and a. In Endo-09's partition matrix equation, the membership values of labelled patterns themselves, denoted by $\bar{u}_{ik}$, are both learning components and balance parameters. In doing so, the reliance on initial membership values of labelled patterns is higher than the other three algorithms.

Summary

Three distance metrics are used by the four algorithms; Fuzzy Mahalanobis distance by Pedrycz-97 and Li-08, kernel-based distance by Zhang-04 and Euclidean distance by Endo-09. Pedrycz-97, Li-08 and Endo-09 algorithms use different approaches to incorporate the supervised learning element into the objective function of the original unsupervised FCM algorithm. This supervised learning element is observed in the second component of Pedrycz-97's and Endo-09's objective functions, and third component in Li-08's. The extended objective functions are then used to derive calculations of centroids and the partition matrix. Rather than using this extension in the objective function, Zhang-04 incorporated supervised learning during the calculation of the centroids. Thus, variations of both unsupervised and supervised learning components feature in these algorithms.

Table 3.5: Datasets used in the experiments. The columns N, n and c specify the number of patterns, features and clusters respectively.

dataset	N	n	c
Iris-2	150	2	3
Iris-4	150	4	3
Wine-2	178	2	3
Wine-13	178	13	3
XOR-2	200	2	4
WOBC-8	699	8	2
WOBC-2	699	2	2
PID-8	768	8	2
PID-2	768	2	2
WDBC-30	569	30	2

3.1.3 Experimental methods

Each algorithm, implemented in R [128], a programming language and environment for statistical computing, was run on ten datasets taken from non-linearly separable datasets Iris, Wine, XOR, Wisconsin Original Breast Cancer (WOBC), Pima Indians Diabetes (PID) and Wisconsin Diagnostic Breast Cancer (WDBC). Note that in Iris, the first class is linearly separable, but the other two are not. The specifications of the datasets are shown in Table 3.5 on page 58. Apart from the XOR dataset which is manually built, all datasets are obtained from the UCI Machine Learning Repository [54]. For each dataset, experiments were repeated with different percentages of labelled patterns, 2%, 4%, 6%, 8%, 10%, 15%, 20%, 25%, 30% and 40% to show the effect of the amount of labelled data has on performance. Using the semi-supervised FCM algorithms, classification can be performed by having the labels replaced with numerals and the labelled patterns ordered such that the clusters and the numerically labelled classes will match. In this way, the unlabelled patterns after clustering can be matched (aligned) with the numerically labelled classes from the datasets before counting the number of correctly matched patterns.

The following settings and modifications were made:

- Zhang *et al.* [165] used initial membership values of 1's and 0's to represent labelled and unlabelled patterns respectively. The assignment

of initial membership values is not mentioned in the other algorithms. However, unlabelled patterns are assumed to hold equal initial memberships of all the clusters, having membership values of $1/c$ (where c is the number of clusters), as was assumed by Bouchachia and Pedrycz [23]. Based on the experiments with 1's and 0's membership in [97], ssFCMs with Fuzzy Mahalanobis tend to run into singularity problem when labelled data are very low due to the many zero membership value in the partition matrix.

- The membership value of each randomly chosen labelled pattern belonging to a cluster is arbitrarily set to 0.9, with 0.1 membership divided among the other clusters. The high membership indicates high belongingness to a cluster. This is applied for the initial partition matrix of all datasets.

- For the Iris, Wine, WOBC and PID datasets, 2-dimensional datasets are created from these datasets to observe the effects of the number of attributes in the datasets on the final clustering outcomes. This was not done for WDBC because arbitrarily reducing a 30-dimensional dataset to 2 dimensions will unlikely improve clustering as the reduced dataset with 2 dimensions arbitrarily chosen will not be informative enough to retain the hidden structure within the data held by 30 dimensions well. Proper feature selection techniques should be employed in this case. As this is a preliminary study, the WDBC dataset was not reduced.

- Dimensions with missing values in the WOBC datasets were removed.

- The boolean matrix in Pedrycz-97 was removed since labelled and unlabelled patterns can be detected from the $\mathbf{F}$ matrix with unlabelled data having $1/c$ membership. In the original ssFCM [121], all data patterns are assigned memberships based on given labels and

stored in **F**. They are then selected to be labelled or unlabelled using the boolean vector **b**. In this case, the labelled data are selected and their memberships are generated prior to running the algorithm. The setting $\mathbf{F} = \mathbf{U^0}$ is kept for the initial partition matrix which contain memberships of labelled and unlabelled data. This setting is applied to all experiments with Pedrycz-97 algorithm.

- The Endo-09 is modified to compute prototypes from the initial partition matrix rather than manually initialising them as the original algorithm did as the membership of labelled pattern in the initial partition matrix will guide the prototype calculation. Manual initialisation may not give the algorithm a good start.

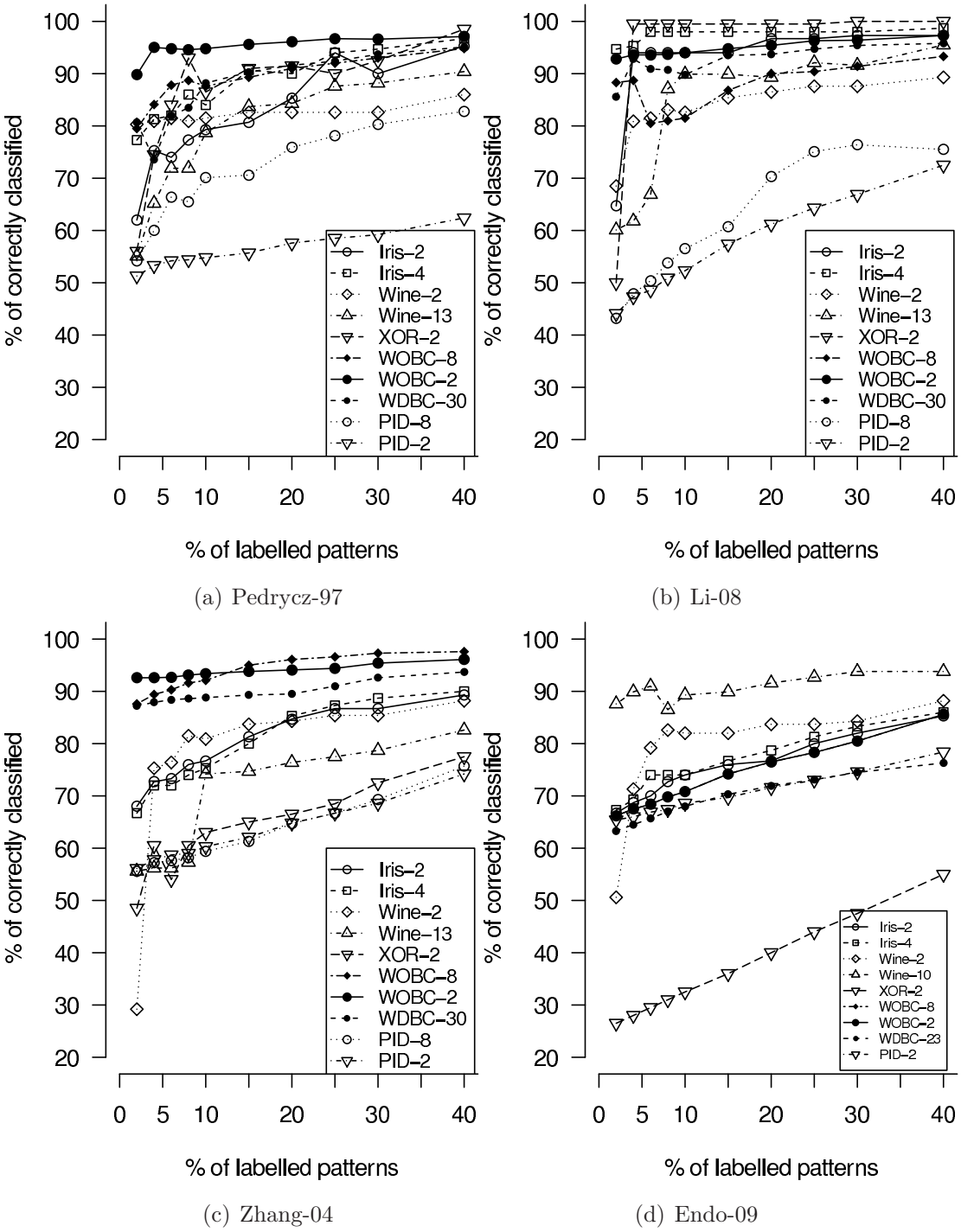
- Endo-09 ran into an infinity problem (the algorithm fails) with datasets Wine-13, PID-8, PID-2 and WDBC-30. These datasets are modified to Wine-10, PID-6, PID-2* and WDBC-23. The modified PID-2 is different from those used by the other algorithms.

- The stopping criterion used in all four algorithms is taken from Pedrycz-97 [121] as follows: $||\mathbf{U'}-\mathbf{U}|| = \sum_{i=1}^{c}\sum_{k=1}^{N}(u'_{ik}-u_{ik})^2$, $||\mathbf{U'}-\mathbf{U}|$ is the sum of squared errors between memberships in previous and current partition matrix U' and U and the threshold value is set to 0.01.

- In Pedrycz-97, scaling parameter α is assumed to be calculated by N/M while a in Li-08 is calculated by $1 - M/N$, where M is the number of labelled patterns and N is the total number of patterns.

- Following the original algorithms, p in Pedrycz-97 and m in Zhang-04 and Li-08 are set to 2 and λ in Endo-09 is set to 1.

- At least one labelled pattern representing each cluster, is required for the clustering to run properly (although semi-supervised algorithms for incomplete labelled patterns already exist [11]).



(a) Pedrycz-97

(b) Li-08

(c) Zhang-04

(d) Endo-09

Figure 3.1: Graph of percentage accuracy against % of labelled patterns.

3.1.4 Results

Figure 3.1 (see Table 7.1 in the appendix on page 195) for more detailed results) show the percentage of correct classifications produced by each algorithm on different datasets. This is a preliminary experiment to get a general understanding of ssFCM, where only one run is conducted for each setting. Overall, Li-08 showed higher accuracy than the other algorithms, achieving more than 80% correct classifications in seven out of ten datasets with 4% of labelled patterns. With 4% of labelled patterns, it produced similar results to Pedrycz-97 in the Wine dataset and Zhang-04 in the WOBC dataset, but outperformed all in the XOR dataset. With 40% labelled data, it produced nearly 100% correct classification in six out of ten datasets. Pedrycz-97 achieved more than 80% correct classification in six out of ten datasets with 6% labelled patterns, while Zhang-04 achieved this with 15% labelled patterns and Endo-09 with 30% labelled patterns. Li-08 produced higher accuracy than Pedrycz-97 in most datasets, but Pedrycz-97 produced higher accuracy than Li-08 in the PID dataset, where the two classes greatly overlap, suggesting that Pedrycz-97 appears better at dealing with clusters with unclear separation.

3.1.5 Discussion

The results of Pedrycz-97 and Li-08 are based on Fuzzy Mahalanobis distance metric, which measures similarity using the sum of squared errors and the correlation with membership values of the patterns. The inverse covariance matrix helps to normalise dimensions of different scales, preventing dominance from dimensions with greater scales. Zhang *et al.* replaced the Euclidean distance metric in FCM with a Gaussian kernel-induced distance metric. The Gaussian kernel measures similarity by $K(x, y) = exp(-||x - y||^2/\sigma^2)$, with the kernel width defined by the variance σ^2. This variance keeps the kernel value of patterns normalised. The idea behind

kernel methods is their ability to solve non-linear problems by mapping the input space into higher dimensional space (the 'kernel trick' [5]), which is applied to distances [137]. Both Fuzzy Mahalanobis and kernel-based distance metrics are competitive, each showing higher accuracy than the other in different datasets with 2% of labelled patterns.

The 'curse of dimensionality' problem is apparent in the results produced by Endo-09 as the algorithm failed on Wine with 13 features and on WDBC with 30 features. The Euclidean distance metric does not reflect scale differences among dimensions in high-dimensional datasets. Dimensions with smaller scales have less influence on the distance in the presence of dimensions with larger scales. This results in distance values that are biased based on scales of the dimensions and this could decrease classification accuracy. Furthermore, with increasing number of dimensions, distance increases exponentially. This makes the Euclidean distance tricky to use in an exponential function, derived from the Endo-09 objective function. Large distance values yield a zero value in the exponential function of Endo-09 partition matrix update in Table 3.4, which in turn returns infinity. Hence, the infinity problem arises in datasets with high dimensions, large scale differences in dimensions and/or large scales in dimensions such as Wine-13, PID-8, PID-6 and WDBC-30. When three dimensions (those with higher scales) are removed from the Wine dataset, the accuracy drastically increased, as shown in Table 7.1 found in the appendix on page 195. Du and Urahama [42] tackled this problem using a parameter δ, which is computed by the inverse sum of average distances with its nearest neighbours and average distances with its furthest neighbours. They showed that their proposed method outperformed semi-supervised kernel methods.

The algorithms did not always produce higher accuracy with datasets that have more dimensions. For example, in Figure 3.1, Pedrycz-97, Li-08 and Zhang-04 produced higher accuracy for WOBC-2 than for WOBC-8.

This is because some features have higher discrimination power than others. Those with low discrimination power add more computational cost and can deteriorate the classification results [85] because they do not provide useful information to discriminate between clusters. Benkhalifa and Bensaid [14] applied feature selection to select subset of features with high discrimination power into ssFCM to categorise text. Through the selection of a subset of features, those that have high discrimination power can be retained while others that deteriorate classification can be discarded.

All algorithms were able to cluster most of the non-linearly separable datasets, achieving more than 80% correct classifications with at most 30% labelled patterns, as shown in Table 7.1. Zhang-04 and Endo-09 achieved less in the XOR dataset and none of the algorithms could achieve more than 80% classification accuracy with 40% labelled patterns in the PID datasets, except for Pedrycz-97. According to Päivinen [117], the two classes in PID are not well-separated, which makes it a challenge for the algorithms.

Interestingly and unexpectedly, increases in the number of labelled patterns did not always increase classification accuracy, indicated by the troughs in the graphs of Figure 3.1. This observation holds for all algorithms except Pedrycz-97 in Wine-2 dataset with 8% to 10% labelled patterns. Overall, there is still a general trend of increasing correct classifications with increase in labelled patterns. This may suggest that not all labelled patterns are good candidates to guide clustering. The choice of suitable labelled patterns, prior to clustering, lies in their features, their initial membership values and the objective function of the algorithm. The presence of labelled patterns that do not strongly belong to one class, but have high membership values will hinder the clustering process, reducing accuracy. The objective function of the algorithm must have some corrective mechanism to handle such patterns, such as those established in [21]. The authors extended the objective function to include the relationship

between classes and clusters, and parameterised the confidence in the accuracy of data labels. The presence of labelled patterns with strong features of a cluster helps give a solid definition of a class. The idea of manually choosing suitable data patterns to be labelled data before they are used to supervise clustering has been previously used in ssFCM learning [65].

Based on the results, Pedrycz-97 and Li-08 appear to perform most favourably. However, this is not definitive as only one run is conducted. Despite the many differences between four algorithms, the results gave a general indication that ssFCMs can perform well on non-linearly separable data, often found in biomedical data. Furthermore, we conclude that the number and scale of dimensions in the data set, distance metrics and objective functions, together, affect clustering results. In addition, not all labelled patterns were found good candidates for supervision.

3.2 Semi-supervised Fuzzy c-Means classification of breast cancer: Investigating distance metrics

3.2.1 Background and motivation

Distance metrics are an important part of Fuzzy c-Means as they are used to measure similarity between data patterns, which provide structural information in terms of the characteristics of data patterns relative to the cluster. The degree of similarity can determine how strongly a data pattern belong to a certain group. Euclidean, Mahalanobis, Fuzzy Mahalanobis and Kernel-based distance metrics use different approaches to measure similarity. They are chosen for investigation as they are popular distance metrics in ssFCM [15, 121, 23, 148]. Hidden structural information can be uncovered using suitable distance metrics that can improve classification results.

Semi-supervised Fuzzy c-Means is used as an automatic technique (post-initialisation) to classify the NTBC dataset. Previously in [99], ssFCM with

Fuzzy Mahalanobis was found to produce poor classification for NTBC. This led to the exploration of different distance metrics to find one that achieves good classification results. Experimental results from using ssFCM with four different distance metrics Euclidean, Mahalanobis, Fuzzy Mahalanobis and Kernel-based are shown. The ssFCM algorithm Pedrycz-97 is used because it has been shown from the work described in the previous section to produce competitive results when compared with Li-08. More importantly, Pedrycz-97 produced higher accuracy than Li-08 when classifying the PID dataset which contain overlapping classes, also common in biomedical datasets. In a separate study investigating initialisation techniques with Li-08 (not included in the thesis), Li-08 was found to be more reliant on the quality of the labels than Pedrycz-97. This is not considered a favourable quality as poor quality labels may affect the results negatively. For these reasons, Pedrycz-97 is chosen for further study.

3.2.2 Experimental methods

The NTBC is classified using ssFCM with Euclidean, Mahalanobis, Fuzzy Mahalanobis and Kernel-based distances, with the purpose to explore how able is ssFCM in finding the the six subgroups identified by Soria *et al.* [141]. Various amounts of labelled data are experimented with; 0%, 10%, 20%, 30%, 40%, 50% and 60% of the 663 classified data patterns. To select data patterns to be labelled, random stratified sampling is applied across the six classes. The experiment is run using each varying amount across 100 different sets of labelled data. The ssFCM setting is based on Pedrycz-97 with the boolean matrix removed as its values are already represented within the $\mathbf{F}$ matrix.

To initialise membership values in $\mathbf{F}$, the selected labelled data patterns belonging to their respective classes are given a membership of 0.9 and (1-0.9)/(6-1)=0.02 for classes they do not belong to. The high 0.9 membership

value is arbitrarily chosen to indicate a data pattern's high possibility of belonging to the class while a 0.02 value indicates otherwise. Unlabelled data patterns have a membership value of $1/c \approx 0.1667$ to indicate equal possibility of belonging to the classes.

To determine the class of a data pattern $\mathbf{x_k}$, the class with the highest membership value is chosen. To evaluate the accuracy of the algorithm, the classes assigned by ssFCM to the 663 data patterns are then compared with Soria's classification [141] and the matches are counted and divided by 663. An average is taken across 100 runs. This is the clustering setting.

The experiments are ran in a classification setting using 10-fold CV where 90% of the 663 data patterns are training data and the remaining 10% is the test data. The algorithm is run 30 times on randomly selected labelled data, across varying amount of labelled data; 0%, 10%, 20%, 30%, 40%, 50% and 60% of training data using the four distance metrics Euclidean, Mahalanobis, Fuzzy Mahalanobis and kernel-based. The classification result obtained from the training process is then used to initialise the algorithm for the testing process. For evaluation, only matches from test labels are counted and divided by the number of test data. An average is subsequently taken across the 30 runs for all 10 folds. This average indicates the agreement level of the solutions with Soria's classification [141] and is presented in terms of percentage. A 100% accuracy result is "optimal" and means the solution completely matches with Soria's classification. Lower accuracy means less matches (similarity) with Soria's classification.

Two evaluation settings have been used because in many ssFCM literatures, experiments were run in a clustering setting, but evaluation conducted in this setting is considered optimistic as it has not been tested on unseen data. Hence, the cross validation technique is also used. For completeness, results based on both evaluation settings are presented.

The solutions are presented using three different evaluation techniques

discussed in Chapter 2.5, accuracy rate expressed in percentage (A), Cohen Kappa Index denoted as κ and Normalised Mutual Index (NMI).

3.2.3 Results

Table 3.6 on page 69 shows the classification results in a clustering setting using ssFCM with Euclidean, Mahalanobis, Fuzzy Mahalanobis and kernel-based distances based on the average percentage of matching class assignments with Soria's classification [141] followed by $\pm$ *standard deviation*. ssFCM with Euclidean distance produced the highest average accuracy, achieving 96.51% accuracy with 10% labelled data. With 50% labelled data or more, almost complete agreement was achieved. This result is expected as the distance metric used in the clustering techniques by Soria *et al.* [141] is the Euclidean distance.

Interestingly, higher accuracy was found using the Mahalanobis distance when compared with the Fuzzy Mahalanobis distance. Fuzzy Mahalanobis was found to produce the worst results in comparison with the other distance metrics on NTBC. To the best of our knowledge, such trend with poorer accuracy using Fuzzy Mahalanobis than Mahalanobis distances has never been reported, despite it being widely used in FCM and ssFCM.

At 0% labelled data, the classification results were very poor as the $1/c$ initial membership was not useful for class discrimination. This caused all data patterns to be assigned to the same class, as the membership remained unchanged after classification. As κ penalises when there is no chance of random agreement with all data patterns assigned to one class, κ becomes zero. NMI fails when the algorithm fails to assign data patterns to all classes, thus the NaN value.

Table 7.2 (on page 196 in the appendix) shows classification results obtained from a classification setting using CV. In a clustering setting, the classification results were observed to be more optimistic than in a CV set-

Table 3.6: Accuracy of ssFCM using Euclidean (E), Mahalanobis(M), Fuzzy Mahalanobis (FM) and kernel-based (K) distances obtained in a clustering setting. The distance metric with highest average accuracy, κ and NMI is indicated in italics, showing that Euclidean with ssFCM is most suitable for NTBC.

DM[1]	ET[2]	0%	10%	20%	30%	40%	50%	60%
	A[3]	30.47±0.46	*96.51±1.32*	*97.74±0.68*	*98.41±0.53*	*98.73±0.43*	*99.05±0.45*	*99.24±0.29*
E	κ [4]	0	0.96±0.02	0.97±0.01	0.98±0.01	0.98±0.01	0.99±0.01	0.99±0.00
	NMI[5]	NaN	0.91±0.02	0.94±0.02	0.95±0.01	0.96±0.01	0.97±0.01	0.98±0.01
	A	30.47±0.46	77.45±2.49	85.89±1.66	89.68±1.21	92.54±0.99	94.24±0.92	95.68±0.74
M	κ	0	0.72±0.03	0.83±0.02	0.87±0.02	0.91±0.01	0.93±0.01	0.95±0.01
	NMI	NaN	0.55±0.04	0.68±0.03	0.75±0.02	0.81±0.02	0.85±0.02	0.88±0.02
	A	30.47±0.46	45.28±3.70	53.66±3.46	60.55±3.26	69.08±3.94	75.65±3.36	82.82±3.23
FM	κ	0	0.29±0.03	0.40±0.04	0.50±0.04	0.61±0.05	0.69±0.04	0.78±0.04
	NMI	NaN	0.30±0.02	0.36±0.03	0.43±0.03	0.52±0.04	0.59±0.04	0.68±0.04
	A	30.47±0.46	73.70±1.84	81.43±1.56	89.03 ±1.20	92.66±0.98	94.85±0.77	96.40±0.55
K	κ	0	0.61±0.02	0.79±0.02	0.84±0.02	0.89±0.01	0.92±0.01	0.94±0.01
	NMI	NaN	0.51±0.03	0.68±0.02	0.73±0.02	0.80±0.02	0.85±0.02	0.89±0.01

[1] Distance Metric
[2] Evaluation Technique
[3] Accuracy in percentage
[4] Cohen's Kappa Index
[5] Normalised Mutual Index

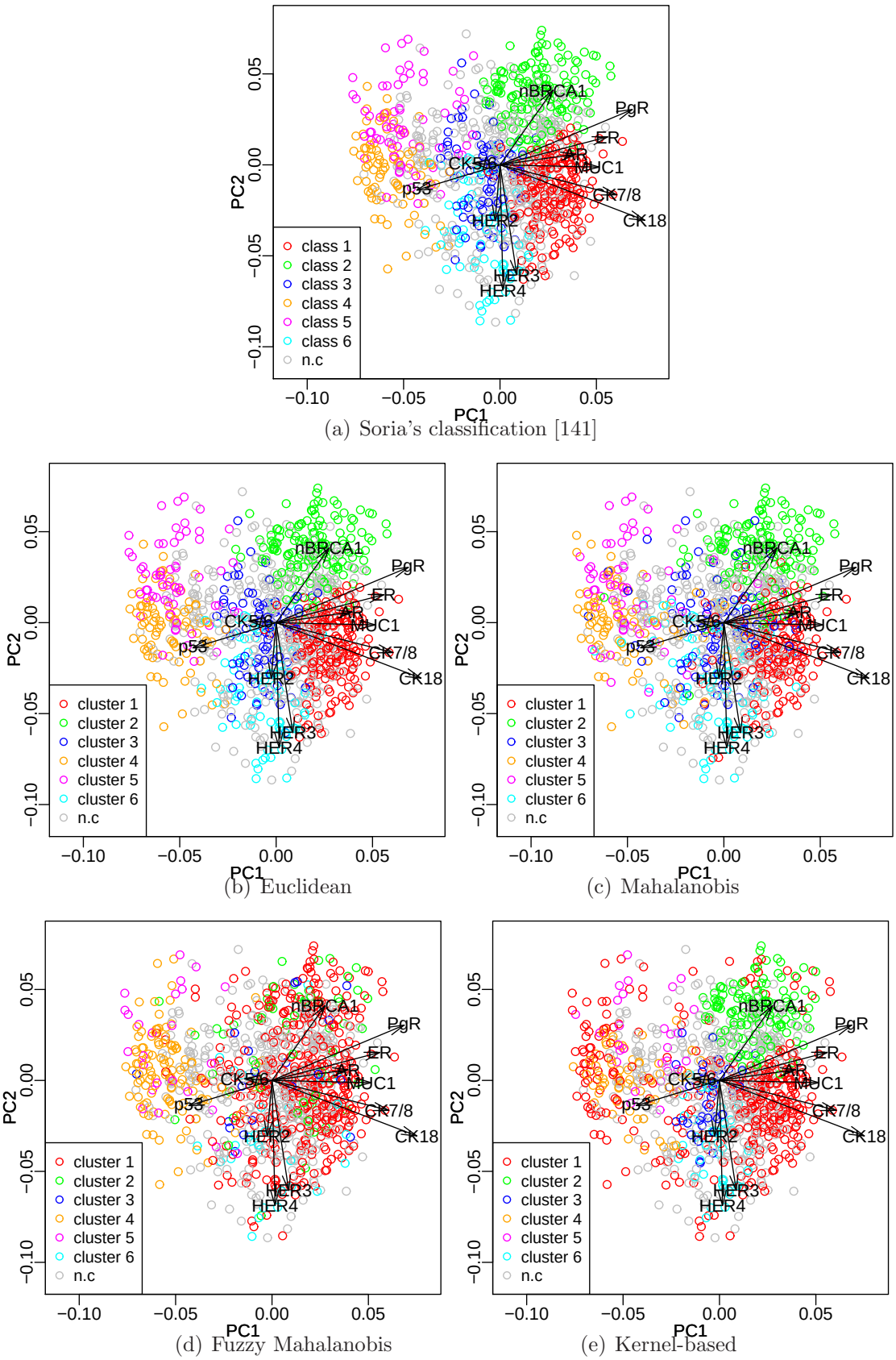
ting. Nevertheless, the trends in the results are similar with ssFCM with Euclidean distance produced the highest average accuracy and ssFCM with Fuzzy Mahalanobis distance the lowest. In the CV setting, ssFCM with Fuzzy Mahalanobis did not assign data patterns to all classes in some runs, resulting in NaN NMI output. These results have therefore been discarded.

Figure 3.2 on page 71 shows biplots of the classification results obtained in a clustering setting. The grey points indicate the 413 not classified (n.c) patients. The classification appears more scattered using the other distances as compared to Euclidean distance. For Fuzzy Mahalanobis distance, class 1 and 4 appear to dominate more than other classes. It is clear from the biplots that the clusters found by Fuzzy Mahalanobis and kernel-based are not those previously identified by Soria *et al.* [141].

3.2.4 Discussion

From the results in Tables 3.6 and 7.2, the choice in distance metrics was demonstrated to greatly impact the accuracy of the classification results. For instance, ssFCM with Euclidean distance produced the highest average accuracy (similarity to Soria's classification) when applied to the NTBC dataset. However, the lowest results were found when employing Fuzzy Mahalanobis distance.

Pedrycz and Waletzky's ssFCM with Fuzzy Mahalanobis distance [121] produced higher accuracy than with the original Mahalanobis distance for UCI Iris dataset and XOR dataset. However, in a separate unpublished study with five UCI datasets (Ionosphere, Page Blocks, Pima Indian Diabetes (PID), Wine and Wisconsin Original Breast Cancer (WOBC)), Fuzzy Mahalanobis distance was found to perform less favourably than Mahalanobis distance for PID, Wine and WOBC datasets, suggesting that Fuzzy Mahalanobis distance does not always produce favourable results than the original Mahalanobis distance for all datasets.



(a) Soria's classification [141]

(b) Euclidean

(c) Mahalanobis

(d) Fuzzy Mahalanobis

(e) Kernel-based

Figure 3.2: Accuracy of various distance metrics on ssFCM with 10% labelled data in a clustering setting.

In another unpublished study, the same experiments were rerun with the five UCI datasets normalised. Euclidean distance was expected to produce the highest accuracy since the datasets were normalised to the same scale for all datasets. Different distance metrics with ssFCM, however, were found to produce best results for different datasets. It is, thus, crucial to experiment with several distance metrics to find one that best represent the dataset, be it normalised or not. Furthermore, Duda *et al.* [43] warned of the dangers of imposing a fixed structure instead of finding it when making a choice on distance metrics.

Without labelled data, the results were significantly poorer. ssFCM assigned all data patterns to one class. The significantly poor results without labelled data was consistent with the findings by Soria *et al.* [141], where the authors reported that poor clustering results were obtained using FCM (an equivalent of ssFCM without labelled data). The significant improvement by using just 10% labelled data is a strong indication that semi-supervised learning can play an important role, even when only a small percentage of the data is labelled. In another study [99], FCM using random initialisation and Euclidean distance (with manual class assignments via visual inspection) achieved only 67.97% accuracy as compared to 96.51% using ssFCM with 10% labelled data. The $1/c$ membership initialisation for the entire partition matrix in experiments with 0% labelled data in [100] provided a very inaccurate contribution from the data patterns into computing the cluster centres. In addition, the clusters found using 0% labelled data are not matched (aligned) to Soria's classification [141] automatically, which is another reason for poor results. The clusters found have to be manually aligned with Soria's classes before matching them with his classification to calculate clusters' agreement.

NMI fails when at least one class does not contain any patterns. As there was limited test data (of 66 or 67 patients) to be classified in six

classes, the disagreements become more exaggerated using NMI and κ. Results from ssFCM with Euclidean distance evaluated using accuracy, κ and NMI are not exceedingly different. This study confirmed the use of Euclidean distance in ssFCM for the proceeding investigations as it produced the highest average accuracy for NTBC. Moreover, the most suitable algorithm can be determined using any of the three evaluation techniques. For these reasons, investigations on classification NTBC are evaluated based on accuracy in a 10-fold CV setting unless stated otherwise.

3.3 Comparisons of ssFCM with other classifiers for breast cancer classification

3.3.1 Background and motivation

Previously, ssFCM with Euclidean distance was shown to produce accuracy of above 90% using 10% labelled data for the NTBC dataset. In this section, the objective is to determine whether ssFCM is a good technique for classifying the NTBC dataset. To do this, popular supervised and semi-supervised learning techniques are applied for classifying the NTBC dataset and their results are compared to ascertain how relatively well the techniques preserve the main clinical groups and ideally identify the same subgroups as those identified by Soria *et al.* In so doing, ssFCM can be demonstrated as a good choice for classifying the NTBC dataset and further investigations in line with our research objectives can be carried out.

3.3.2 Selected algorithms

In this section, the algorithms selected for comparison are briefly reviewed. Due to their popularity or their characteristics suitable for application on the NTBC, these algorithms are selected.

k-Nearest Neighbours

k-Nearest Neighbour (KNN) is one of the most fundamental and simple classification methods. For this reason, it should be one of the first choices of classification methods [125]. Its non-parametric nature means that it makes no assumption about the data distribution. It classifies objects based on majority vote of its closest data patterns (neighbours), where k is the number of nearest neighbours. The neighbours are determined by a chosen distance metric, which defines the similarity measure. For a query data pattern $\mathbf{x_q}$ to be classified, the nearest k neighbours of $\mathbf{x_q}$ are computed and the class represented by the majority of the neighbours is returned. Care has to be taken in choosing a suitable k value. Larger values of k reduce the effect of noise, but can make class boundaries less distinct.

C5.0

C5.0 is an improved version of C4.5 [126], both of which generate decision trees for classification using tree induction methods. While the improvements in C5.0 are documented in [127], the details of the extension remain largely undocumented. Thus, the overall idea which is based on the C4.5 algorithm is explained instead.

Given a set of training data, $\mathbf{S} = \mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_n}$ where each sample $\mathbf{s_i}$ consist of p-dimensional vector $(x_{1,i}, x_{2,i}, ..., x_{p,i})$, $\mathbf{x_j}$ represents features of the sample as well as which class $\mathbf{s_i}$ belongs to. Based on the normalised information gain as a splitting criterion, the most effective attribute at each node of the tree is chosen to split the data into subsets which represent the classes. The algorithm is recursive and moves on to smaller subsets.

Random Forests

Random forests [27] construct a combination of classification (decision) trees during training. To classify a new data pattern, an input vector is

classified by each tree and the class with the most votes is selected. Each tree is constructed as follows:

1. Let the number of training cases be N and the number of variables be M. The training set is selected by choosing n times with replacement from all N training cases, i.e choosing any training case from the population more than once.

2. At each node, m variables out of M are randomly selected such that $m < M$. To split the node, the best split based on the m variables is used. The best split can be calculated using an impurity measure such as information gain.

3. Each tree is fully grown and no pruning is performed.

Naive Bayes

The Naive Bayes is a probabilistic classification technique based on Bayes' theorem and assumes independence between features. Consider a supervised learning problem where the aim is to find $f : X \Rightarrow Y$, or equivalently $P(Y|X)$ where X is a vector with features $X_1...X_n$ and Y its corresponding class labels. One way to learn $P(Y|X)$ is to use labelled data to estimate $P(X|Y)$ and $P(Y)$. These estimates, with Bayes rule, can be used to find $P(Y|X = \mathbf{x_k})$ for a new data pattern $\mathbf{x_k}$. The aim is to train a classifier to output the probability distribution over possible values of Y for each new data pattern X. Assuming that X_i are conditionally independent given Y, and Y takes on its kth possible value, the fundamental equation for Naive Bayes is expressed as:

$$P(Y = y_k|X_1...X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)} \tag{3.1}$$

where X has n conditionally independent features given Y, expressed as:

$$P(X_1...X_n|Y) = \prod_{i=1}^{n} P(X_i|Y) \tag{3.2}$$

Generalised Linear Models with Elastic Nets

Friedman *et al.* [55] developed fast algorithms to estimate generalised linear models including multinomial regression problems with convex penalties such as the lasso (ℓ_1) [151] and ridge regression (ℓ_2) [80]. These Generalised Linear Models with Elastic Nets (GLMNET) algorithms use cyclical coordinate descent, which is computed along a regularisation path. The idea of GLMNET is to solve the lasso problem using coordinate descent. This is done by optimising each parameters separately and holding all of the rest fixed. This procedure is cycled until the coefficients stabilise.

Given a response variable (class labels) $Y \in \mathbb{R}$ and a predictor vector $\mathbf{X} \in \mathbb{R}^p$, the regression function is approximated using a linear model $E(Y|X = x) = \beta_0 + x^T\beta$, there are N observation pairs $(\mathbf{x}_i, y_i)$ and $\mathbf{x}_i$ is a vector containing x_{ij} for $j = 1, ..., p$. The elastic net solves the following problem:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 + \lambda P_\alpha(\beta) \right], \tag{3.3}$$

where

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}||\beta||_{\ell_2}^2 + \alpha||\beta||_{\ell_1} \tag{3.4}$$

$$= \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right] \tag{3.5}$$

To solve (3.3), coordinate descent step is used such that estimates $\tilde{\beta}_0$ and $\tilde{\beta}_l$ for $l \neq j$ and (3.3) are partially optimised with respect to β_j.

Neural Networks

Neural networks (refers to artificial neural networks) are generally biologically-inspired, adaptive systems which change their structure based on inputs

and connections of the networks. They are made up of inter-connections between neurons located in different layers of each system. Feed-forward neural networks with a single hidden layer [133, 134], for instance, aim to learn the weights of interconnections w_i so that $f(x) = K(\sum_i w_i g_i(x))$ where K is the activation function which compute a neuron's output activation based on its weighted input and $g_i(x)$ are a collection of functions. By learning the weights, the algorithm can find a function $f : X \rightarrow Y$ where X is a data pattern represented by a vector and Y its class label.

Learning Vector Quantisation

Learning Vector Quantisation (LVQ) [93] is a supervised form of vector quantisation and is a prototype-based classification algorithm. Built using a self organizing map [92], with a winner-take-all approach, it aims to find a set of prototypes (weights), $\mathbf{w_j}$ that best represent each class using training input vectors, $\mathbf{x}$. The idea is to shift the Voronoi cell boundaries to achieve better classification. The weight vector $\mathbf{w_j}$ is updated by checking the input classes against the Voronoi cell classes as follows:

1. If input $\mathbf{x}$ and weight $\mathbf{w_{I(x)}}$ (where $j = I(\mathbf{x})$ is the winning output neuron) have the same class label, then update the new weight vector as $\mathbf{w_{I(x)}}' = \mathbf{w_{I(x)}} + \alpha(\mathbf{x} - \mathbf{w_{I(x)}})$ to move $\mathbf{w}$ towards $\mathbf{x}$. The class labels of the output neurons are preassigned. The winning output neuron is the one with weight vector $\mathbf{w_{j=I(x)}}$ closest to $\mathbf{x}$, where the distance used is Euclidean.

2. If $\mathbf{x}$ and $\mathbf{w_{I(x)}}$ have different class labels, $\mathbf{w}$ is moved away from $\mathbf{x}$ such that $\mathbf{w'_{I(x)}} = \mathbf{w_{I(x)}} - \alpha(\mathbf{x} - \mathbf{w_{I(x)}})$,

where α is a learning rate that decreases at each iteration. The algorithm stops when the learning rate reaches a threshold value.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is often referred to as the Fisher's linear discriminant [52]. Its objective is to maximise class separability by maximising the difference between the means, normalised by measure of within-class scatter. The separation between the two distributions is defined as the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(w \cdot \mu_2 - w \cdot \mu_1)^2}{w^T \Sigma_2 w + w^T \Sigma_1 w} = \frac{w \cdot (\mu_2 - \mu_1))^2}{w^T (\Sigma_1 + \Sigma_2) w} \tag{3.6}$$

where μ_1 and μ_2 are the means of the two classes and their covariances are Σ_1 and Σ_2. The linear combination of features $w \cdot x$ has means $w \cdot \mu_i$ and covariances $w^T \cdot \Sigma_i w$ for $i = 1, 2$. The maximum separation can be shown to occur when $w = (\Sigma_1 + \Sigma_2)^{-1} (\mu_2 - \mu_1)$.

Mixture Discriminant Analysis

Mixture Discriminant Analysis (MDA) [74] is a classification method based on mixture models. It is an extension of LDA, aimed at improving LDA's restriction of linear boundaries which models a class using a single Gaussian. Instead, a class can be represented using a mixture of Gaussians and non-linear boundaries. The overall model is expressed as:

$$P(X, G = j) = \sum_{r=1}^{R} \pi_r \phi(X; \mu_r, \Sigma) P_r(j). \tag{3.7}$$

The model is a mixture of joint-densities $P_r(X, G)$ with R shared mixture components where the rth mixture density has prior probability π_r for class j such that $\sum_{r=1}^{R} \pi_r = 1$ and $P_r(j)$ is the prior probability of class j. ϕ is the multivariate Gaussian density function with parameters μ_r as its mean and Σ as its covariance. The EM algorithm is used to estimate π_r, μ_r and Σ.

High Dimensional Discriminant Analysis

The High Dimensional Discriminant Analysis (HDDA) [24] estimates specific subspaces within the data and the intrinsic dimension of these classes. Thus, HDDA reduces the dimension for each class independently and performs regularisation of class conditional covariance matrices to adapt the Gaussian framework to high dimensional data. HDDA is based on the assumption that high dimensional data exist in different subspaces with low dimensionality. The idea is to work in class subspaces with lower dimensionality, assuming the classes are spherical in these subspaces.

Kernel Support Vector Machines

The optimal hyperplane algorithm was originally a linear classifier introduced by Vapnik [36]. To create non-linear classifiers, Boser *et al.* [20] applied the "kernel trick" [5] to maximum-margin hyperplanes. The dot product $\langle \mathbf{z_i}, \mathbf{x} \rangle$ is replaced by a non-linear kernel function resulting in the output for Kernel Support Vector Machines (KSVM) as follows:

$$F(x) = \sum_{i=1}^{N} w_i k(\mathbf{z_i}, \mathbf{x}) + b \tag{3.8}$$

where $\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_N}$ are support vectors and $w_1, w_2, ..., w_N$ are weights. The kernel trick allows kernel functions to map input vectors to a higher dimensional space, useful for solving non-linear problems.

Transductive Support Vector Machines

The transductive SVM (TSVM) [87] is an extended technique of SVM. TSVM depends on the construction of the classification hyper-plane by inductive learning on labelled training samples, and getting the discriminate function value for each unlabelled samples, following the discriminate function, $f(\mathbf{x}) = \mathbf{w0} \cdot \mathbf{x} + b0$. By using current labelled samples, TSVM gets the current split-plane and the discriminate function described previously, and

calculates all current discriminate function values of the unlabelled samples [31]. The 2-norm for TSVM is methodically adopted. The standard setting can be illustrated as:

$$
\begin{aligned}
&\text{Minimize over } (\mathcal{Y}_1^*, ..., \mathcal{Y}_k^*, \mathbf{w}, b, \xi_1, ..., \xi_m, \xi_1^*) \\
&\frac{1}{2}||\mathbf{w}||_2^2 + \mathbf{C}\sum_{i=1}^{m}\xi_i + \mathbf{C}^*\sum_{j=1}^{k}\xi_j^* \\
&\text{Subject to:} \\
&\forall_{i=1}^{m} : \mathcal{Y}_i\left(\mathbf{w^T}\emptyset\left(\mathbf{z} * \mathbf{x_i}\right)' + b\right) \geqslant 1 - \xi_i, \xi_i \geqslant 0 \\
&\forall_{j=1}^{k} : \mathcal{Y}_j^*\left(\mathbf{w^T}\emptyset\left(\mathbf{z} * \mathbf{x_j^*}\right)' + b\right) \geqslant 1 - \xi_j^*, \xi_j^* \geqslant 0
\end{aligned} \tag{3.9}
$$

where $\mathcal{Y}_j^*$ is the unknown label for $\mathbf{x_j^*}$ which is one of the k unlabelled samples, (3.9) takes the unlabelled data into consideration, by representing the violation terms ξ_j^* caused by forecasting each unlabelled pattern $\emptyset\left(\mathbf{x_j^*}\right)$ into $\mathcal{Y}_j^*$. The penalty to these violation terms is controlled by new constant $\mathbf{C}^*$ with unlabelled samples while $\mathbf{C}$ consist of labelled samples only [86, 30].

Precisely solving the transductive problem involves searching all potential assignments of $\mathcal{Y}_1^*, ..., \mathcal{Y}_k^*$ and identifying various terms of ξ^* which is regularly intractable for big data sets.

3.3.3 Experimental methods

Using 10-fold CV with 90% training data and 10% test data, the highest results from ssFCM (with the most suitable distance metric) are compared with other classifiers GLMNET, C5.0, LDA, HDDA, KNN, NNET, NB, MDA, KSVM, RF, LVQ and TSVM. For TSVM, the experimental set-up is similar to ssFCM where various amount of labelled data, 10% to 90% of labelled data are chosen from training data.

TSVM is implemented in the SVMLight [4, 58]. The `train` function in the `caret` R package [96] is used to implement the other classification

Table 3.7: Accuracy comparison of Euclidean ssFCM (with 10% and 90% labelled data) with other classifiers.

ssFCM10	ssFCM90	GLMNET	C5.0	LDA
96.12±2.04	97.80±1.49	96.50±3.07	86.42±6.06	95.17±1.57
HDDA	KNN	NNET	NB	MDA
93.97±1.72	94.58±3.02	59.90±11.69	80.56±5.95	94.27±2.84
KSVM	RF	LVQ	TSVM10[79]	TSVM90[79]
94.29±3.06	95.79±2.51	95.18±2.33	88.09±2.88	99.61±0.59

techniques, with default parameter settings based on the package. Thus, no tuning parameters have been set apart from the number of classes. The NNET used is a feed-forward single-hidden layer neural network.

A 100% accuracy result is "optimal" and means the solution completely matches with Soria's classification. Lower accuracy means less matches (similarity) with Soria's classification.

3.3.4 Results

Table 3.7 shows the accuracy of classification using the selected techniques previously described. Apart from C5.0, NNET, NB and TSVM with 10% labelled data, the other techniques performed competitively well on NTBC, with accuracy of above 90%. GLMNET, LDA, RF and TSVM with 60% labelled data, in particular, have achieved above 95% accuracy. ssFCM with only 10% of labelled data produced one of highest matching results to Soria's classification.

Table 3.8 shows the accuracy comparison between Euclidean ssFCM and TSVM using 10% labelled data to 90% labelled data. ssFCM was observed to produce higher accuracy than TSVM at low amounts of labelled data between 10% to 40%. For some runs, ssFCM was able to achieve a maximum accuracy of 100% even with only 10% labelled data. However, at 75% to 90% of labelled data, TSVM outperformed ssFCM. At 90% of labelled data, TSVM achieved an average accuracy of 99.61% while ssFCM achieved an average accuracy of 97.80%.

Table 3.8: Accuracy comparison between Euclidean ssFCM and TSVM [79].

ssFCM	10	20	30	40	50	60	70	80	90
average	96.12	96.86	97.22	97.54	97.43	97.84	97.81	97.85	97.80
st.dev	2.05	1.94	1.78	1.62	1.64	1.53	1.53	1.52	1.49
TSVM [79]	10	20	30	40	50	60	70	80	90
average	88.09	91.33	96.01	96.41	97.36	97.71	98.06	98.90	99.61
st.dev	2.88	2.81	3.14	0.69	1.31	1.67	1.51	0.90	0.59

3.3.5 Discussion

Mitchell [114] explained that when using NB for continuous inputs, the variable inputs are assumed to follow a Gaussian distribution. Due to this assumption, NB was found to performed poorly on the NTBC dataset as the feature values are highly non-normal. Interestingly, both MDA and KSVM under non-linear settings have been slightly outperformed by LDA with linear boundaries restriction. Perhaps, LDA had found good compromised boundaries to provide a more generalised classification (less overfitting) for the dataset as compared to more exacting boundaries of MDA and KSVM.

TSVM (with 60% of labelled data) outperformed KSVM (with 90% labelled data) despite TSVM having lesser labelled data than KSVM, showing that a semi-supervised approach is favoured for classifying NTBC. Furthermore, TSVM produces non-linear boundaries like MDA and KSVM. C5.0 did not perform as well in comparison with the other classifiers studied. However, decision-tree-based RF performed well on the NTBC. This suggests that having a voting system within a forest of decision trees such as in RF improved classification for decision-tree-based classifiers. The single hidden layer, the size of the hidden layer and random initialisations in NNET may have caused the poor classification. Furthermore, default settings were used without any regularisation, which could explain the poor performance. More tuning and experimentation are therefore required to

optimise the solution. On the contrary, LVQ, a special case of neural networks, is guided by prototypes in the competitive layer and this has produced high accuracy.

A comparison between ssFCM and TSVM showed that TSVM achieved higher average accuracy at an almost completely supervised setting since TSVM outperformed ssFCM when using 70% or more labelled data. ssFCM, on the contrary, outperformed TSVM significantly using only 10% labelled data. The two techniques have the advantage at two different situations, ssFCM when availability of labelled data is low and TSVM when availability is high. The Euclidean distance in ssFCM may have confined data into hyperspherical clusters whose shapes stabilised at 60% labelled data while TSVM continued to evolve the non-linear hyperplanes to produce good margins to separate between classes using more labelled data.

3.4 Summary

As preliminary investigative work in ssFCM, this study was aimed at investigating some existing distance-based semi-supervised Fuzzy c-Means algorithms running on popular datasets to determine the common issues that affect the clustering performance in ssFCM. The classification results of four such algorithms, i.e, Pedrycz-97, Li-08, Zhang-04 and Endo-09, with $1/c$ initial membership value of unlabelled patterns, were compared. Based on experimental results, issues such as scale differences of dimensions in the data set, distance metrics, objective functions and quality of labelled patterns were found to affect the classification results. Furtheremore, Fuzzy Mahalanobis distance was observed to produce a more favourable performance than Gaussian kernel-based distance, and that the Euclidean distance metric performed least well, on the selected data sets.

Despite arbitrarily selected features for experimentation, some algorithms achieved higher accuracy on dataset with lesser features than the

original number. An investigation using systematic feature selection with ssFCM can therefore help determine if classification results can be improved with less number of features. This is useful because the removal of features with less discriminating power may actually improve classification and reduce data collection and processing time with less number of features.

Algorithms that can achieve very high percentage of correct classifications with few labelled patterns are desirable. Also, the percentage of correct classifications should increase with increasing percentage of labelled patterns since more examples are available to guide the clustering. It was observed that increases in the number of labelled patterns did not always increase classification accuracy, which suggests that not all labelled patterns are good candidates to guide clustering. Some labelled patterns with poor discriminating power, when used in some algorithms, can negatively impact on classification results. The choice of suitable labelled patterns, prior to clustering, lies in their features, their initial membership values and the objective function of the algorithm.

Using Euclidean ssFCM, accuracy of near 100% could be achieved using only 10% labelled data on the NTBC dataset. Thus, the same breast cancer classes as those previously found by Soria *et al.* have been successfully identified and with a high level of accuracy using ssFCM with Euclidean distance. The ssFCM with Fuzzy Mahalanobis distance was found not to always produce the most favourable results for some datasets. More importantly, the accuracy can be greatly improved when a suitable distance metric is used. It is thus crucial to investigate the various distance metrics to identify the distance metric best suited for the dataset.

The NTBC dataset was classified using various popular classifiers and the classification accuracies were compared between them and ssFCM. ssFCM was found to produce one of the highest results (indicating high similarity to Soria's classification), even when using only 10% labelled data,

making it a suitable technique for classifying the NTBC dataset. Other classifiers such as RF, GLMNET, LVQ and LDA have also produced competitive results. It was observed that classification accuracy stabilised at 60% labelled data for ssFCM while it continued to increase for TSVM with more labelled data, indicating that ssFCM performs most favourably where labelled data are scarce while TSVM reaches near optimal performance at near completely supervised setting for the NTBC.

- This page is empty -

4 Approaches For Improving ssFCM

In this chapter, three approaches for improving ssFCM are investigated; initialisation, feature selection and adjustment of scaling parameter α.

4.1 Initialisation techniques

4.1.1 Background and motivation

Fuzzy clustering algorithms are initialised by either supplying initial membership values or initial cluster centres (step 1 or 2 respectively of ssFCM algorithm in Chapter 2.4.1). The initialisation of membership values involves the assignment of membership values to data patterns. Whereas, the initialisation of cluster centres involves assigning vectors to be representatives of clusters. Traditionally, a number of clustering runs are conducted with different sets of randomly initialised cluster centres or membership values. Cluster validity indices such as Dunn index [45] or Davies-Bouldin index [40] are then used to evaluate the clusters from these runs. The final cluster centres or membership values that produce the best evaluation scores are selected as the clustering solution. As the results produced by the fuzzy clustering algorithms vary with different initialisations, they are regarded as sensitive to initialisation.

In ssFCM, membership values can be initialised using available labelled data patterns instead of using random initial membership values. Sometimes, poor initial membership values may be introduced which can negatively affect classification results. Initialisation techniques have been proposed to give clustering algorithms a good start and to produce favourable results [33]. Yager and Filev proposed the mountain method [164], which finds initial prototypes based on potential values of grid points. The po-

tential values are calculated from grid points to data pattern distances. As potential values of every grid points have to be calculated, the mountain method is known to be computationally expensive. To counter this, Chiu [32] proposed a prototype estimation technique based on potential values of data patterns within a specified radius. While much of the initialisation research in clustering have been focused on K-means [25, 78], little literature is found on initialisation techniques in ssFCM algorithms.

In this section, the effect of initialisation techniques on classification accuracy of the ssFCM algorithm proposed by Pedrycz and Waletzky [121] is investigated. By using initialisation techniques, the unsupervised-generated cluster centres, in addition to labelled data, can be used as additional supervision to improve classification accuracy. Application of initialisation techniques in Fuzzy clustering using cluster estimation (CE) by Chiu [32] was conducted by Liu *et al.* [104] and have produced favourable results. However, this has not been applied in ssFCM algorithms. The objectives of this study are of two-folds. Firstly, to investigate the effect of initialisation techniques on ssFCM classification of the UCI datasets and secondly, to apply this approach on real-world biomedical data, the NTBC dataset, to improve ssFCM classification results.

4.1.2 Experimental methods

Prior to running ssFCM, an initialisation technique is used to find the initial cluster centres $\mathbf{V^0}$ for the entire dataset as illustrated on Figure 4.1. $\mathbf{F}$ is the supervision matrix, $\mathbf{U^0}$ is the initial membership matrix, $\mathbf{U}'$ is the final membership matrix (partition matrix) and $\mathbf{U^{UL}}$ is the membership matrix of unlabelled data. The initial clusters replace step 2 of the ssFCM algorithm for the first iteration on page 28. This means that instead of using the labelled data to calculate the initial cluster centre as in (2.20) on page 28, an initialisation technique that is external of ssFCM is used on

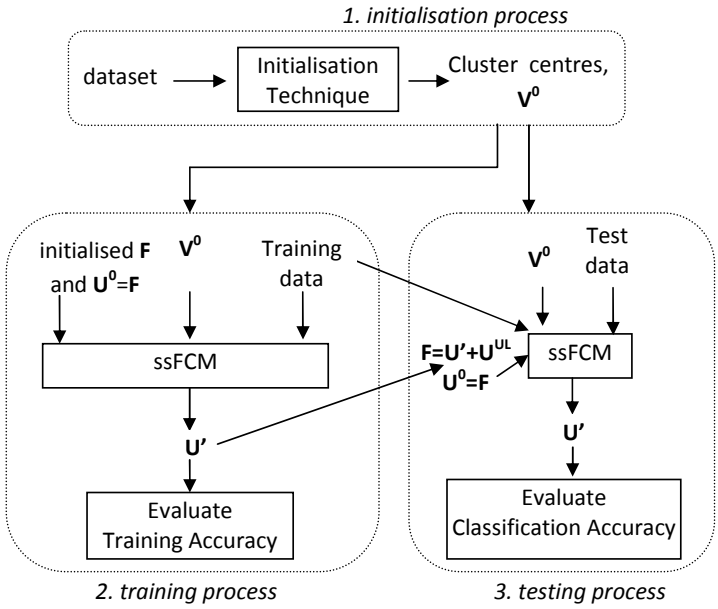


Figure 4.1: ssFCM framework with initialisation techniques

the entire dataset to find the initial cluster centres, to give a good start to the algorithm, aiming to improve accuracy. The initialisation techniques selected for investigation are CE, SCS and KKZ (described in Chapter 2.6).

As a preliminary study, three different initialisation techniques with ssFCM are explored on three popular UCI datasets, Iris, Wine and Pima Indian Diabetes (PID) in a clustering setting. The specifications for these datasets can be found in Table 3.5 on page 58. Varying amounts of labelled data (10%, 20%,30%,40%, 50% and 60% of the total number of data patterns) are used. Labelled data are chosen randomly using stratified sampling and 30 runs are conducted. The three initialisation techniques with ssFCM are subsequently applied on the NTBC dataset using 10-fold CV where varying amounts of labelled data are chosen from the training data. The framework using cross-validation is illustrated in Figure 4.1. The cluster centres generated by the initialisation technique are used in both the training and testing stage. Accuracy is expressed in percentage of the amount of matching class labels over total number labels. For NTBC, a 100% accuracy result is "optimal" and means the solution completely matches with Soria's classification. Lower accuracy means less matches (similarity) with Soria's classification.

The initial cluster centres and final cluster centres after ssFCM converges are presented on a 2-dimensional biplot for further analysis. The biplot is based on the first two principal components obtained from running Principal Component Analysis (PCA) on the dataset. No PCA components are used in the actual classification.

4.1.3 Results

From Table 4.1, some of the initialisation techniques were found to increase the average accuracy of ssFCM on the UCI datasets. For Iris, results of ssFCM with KKZ and CE shows increase in average accuracy when compared with ssFCM. For Wine, results of ssFCM with KKZ shows increase in average accuracy when compared with ssFCM. For PID, average accuracy increased using SCS with ssFCM. Increase in average accuracy was found particularly when using initialisation technique with ssFCM on datasets with low amount of labelled data at 10%.

Table 4.2 shows the classification result of ssFCM using Euclidean distance with initialisation techniques SCS, KKZ and CE on NTBC. Using KKZ, ssFCM was found to produced higher average accuracy results, especially when availability of labelled data is low, as shown in Table 4.2. At higher availability of labelled data, although the average accuracy increased slightly, this slight increase in average accuracy is still considered important for prediction of breast cancer subgroups.

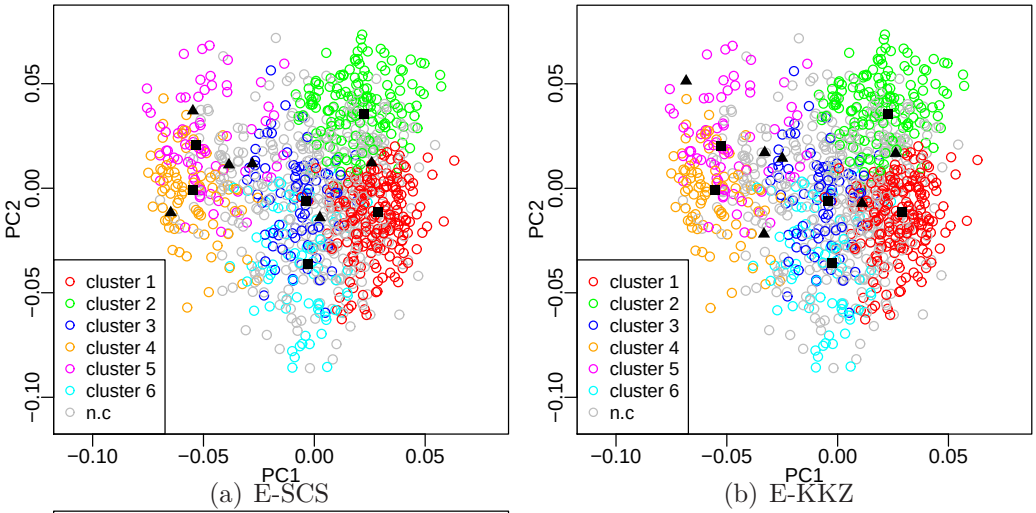
Figure 4.2 on page 92 shows the (cluster) centres generated by SCS, KKZ and CE denoted by ▲ and the centres after ssFCM iterations denoted by ■. KKZ appears to generate centres at the edge of the clusters. The centres by SCS appear to mostly locate on the left side of the biplot. This was also observed with CE. Despite the centres having a larger concentration on the left side of the biplot, ssFCM was able to converge with the final centres near the actual centres.

Table 4.1: Accuracy of Fuzzy Mahalanobis ssFCM and initialisation techniques SCS, KKZ and CE on the UCI datasets. Where there is increase in average accuracy using initialisation techniques, the results are indicated in italics.

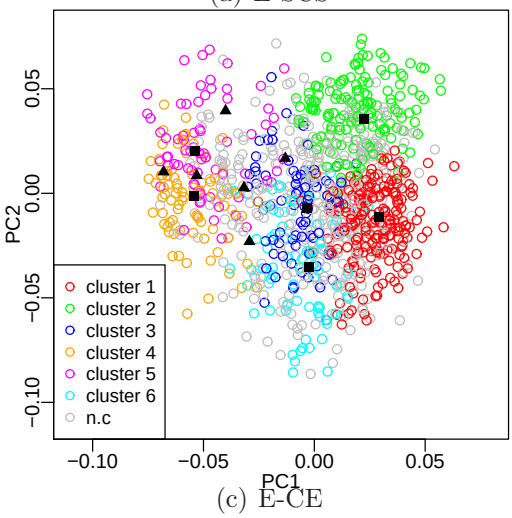
	10%	20%	30%	40%	50%	60%
Iris						
FM-ssFCM	90.98±9.51	95.24±2.91	97.09±1.19	98.20±0.86	98.42±0.90	98.96±0.52
FM-SCS	*92.40±7.34*	*95.47±2.04*	96.73±1.68	97.96±1.06	98.20±1.12	98.87±0.56
FM-KKZ	*92.42±11.31*	*97.36±0.73*	*97.93±0.75*	*98.51±0.62*	*98.69±0.73*	*99.00±0.49*
FM-CE	*95.09±2.58*	*95.78±2.19*	*97.13±1.30*	*98.24±0.88*	*98.47±0.91*	*98.98±0.52*
Wine						
FM-ssFCM	88.15±5.45	92.64±3.73	94.57±1.70	96.14±1.42	97.06±1.09	98.03±0.96
FM-SCS	*90.88±3.16*	*93.78±2.26*	*94.83±1.75*	96.12±1.47	97.10±1.03	*98.11±0.96*
FM-KKZ	*89.29±5.51*	*93.46±3.15*	*95.02±1.72*	*96.46±1.29*	*97.28±1.02*	*98.26±0.91*
FM-CE	*88.37±4.74*	*92.77±3.62*	*94.98±1.89*	96.10±2.10	*97.28±1.03*	98.01±0.99
Pima Indian Diabetes						
FM-ssFCM	62.90±5.96	71.86±2.83	77.87±2.33	82.63±1.68	86.22±0.96	89.18±0.86
FM-SCS	*63.73±5.73*	*72.04±2.48*	77.64±2.51	*82.72±1.63*	*86.26±0.93*	*89.32±0.87*
FM-KKZ	59.54±6.33	69.61±3.88	75.83±2.16	81.68±1.69	86.04±0.98	88.75±1.12
FM-CE	62.46±4.76	71.68±3.85	76.93±2.14	82.01±1.49	86.04±0.96	88.90±0.89

Table 4.2: Accuracy of ssFCM using Euclidean (E) distance and initialisation techniques SCS, KKZ and CE on NTBC. Where there is increase in average accuracy using initialisation techniques, the results are indicated in italics.

Dist.	10%	20%	30%	40%	50%	60%
E	96.12±2.04	96.86±1.94	97.22±1.77	97.54±1.61	97.64±1.55	97.84±1.53
E-SCS	*96.38±1.96*	96.94±1.85	97.24±1.76	*97.55±1.58*	97.61±1.52	97.77±1.54
E-KKZ	*96.50±1.95*	*97.06±1.86*	*97.43±1.72*	*97.63±1.57*	*97.73±1.50*	*97.85±1.51*
E-CE	95.87±2.04	96.69±1.94	97.04±1.81	97.40±1.64	97.46±1.60	97.77±1.59



(a) E-SCS (b) E-KKZ



(c) E-CE

Figure 4.2: Cluster centres generated by initialisation techniques denoted by triangles (▲) and found by initialisation techniques with ssFCM (with 10% labelled data) denoted by squares (■). The coloured data patterns are based on Soria's classification to show where the clusters are.

4.1.4 Discussion

Using ssFCM with initialisation techniques, an increase in average accuracy was found on the UCI datasets and on NTBC. In general, increase in average accuracy was found where amount of labelled data is less, which suggests that initialisation techniques can help increase ssFCM accuracy where labelled data are scarce. This is important as the initialisation techniques are able to recommend data patterns that are potentially good representatives of the clusters when labelled data are few. Although the increase in accuracy may be considered slight, this is considered important especially when the application focus is on biomedical data, such as the NTBC.

The biplot analysis of the cluster centres revealed that KKZ generates

initial clusters at the edge of potential clusters, which is an advantage as they can become cluster prototypes for the nearest clusters. This explains their higher accuracy as compared to SCS and CE for a majority of the datasets. SCS depends on the ordering of data patterns and the ρ threshold to determine cluster centres. If ρ is too large, no cluster centre can be found. On the other hand, if ρ is too small, the cluster centres will be near each other, which explains a higher concentration of cluster centres at the left side of the biplot in Figure 4.2(a). The r parameters in CE becomes tricky to set as "neighbourhoods" are not clearly defined with overlapping clusters and suffers similar problems with SCS with smaller neighbourhoods.

One could argue that the use of initialisation techniques, originally meant for clustering, for classification is not suitable as it carries no logical structure between cluster centre and class labels. Perhaps, the nature of KKZ in identifying initial cluster at the cluster edge makes it ideal to be applied in a classification environment using a clustering algorithm where each cluster represents a class. KKZ gives it that flexibility for a cluster centre to belong to any of the adjacent clusters, which has brought about an increase in average accuracy to the current ssFCM setting for the classification of NTBC.

4.2 Feature selection

4.2.1 Background and motivation

One motivation for this work is to "reproduce" Soria's classification using one clustering method. The long term goal of this chapter is to produce a clinically useful classification using as few features (biomarkers) as possible. Using only important features may not only improve classification accuracy, but also saves time, effort and expense in running clinical tests to obtain the measurable data.

Consequently, the objectives of this work are to three-fold. Firstly, to achieve high classification accuracy using ssFCM and feature selection. Secondly, for clinical interests, the hope is to identify important features for this dataset, reducing the number of features from 25 to a number where increase in ssFCM accuracy can be obtained. Model fitting and stability of selected features are also investigated. Thirdly, to show that improved accuracy can be achieved on other datasets as well, by experimenting on three UCI datasets; Arrhythmia, Cardiotocography and Yeast.

The approach in this chapter is based on a combination of two approaches by Benkhalifa and Bensaid [14] and by Park and Yae [120]. A feature selection technique is first used to select important features. Next, ssFCM is run with the selected features to investigate classification accuracy and evaluate the selected features. The ssFCM-based feature selection methodology is different because the selected features are evaluated based on the highest ssFCM classification results obtained from running ssFCM with varying amount of labelled data, 10% to 60% labelled data. By having this variation in labelled data, the hope is to pick out important features that could achieve high accuracy even with little labelled data.

4.2.2 Experimental methods

The classification methodology is made up of three main processes and is illustrated in Figure 4.3:

1. Feature selection: Labelled data from training set are used by a feature selection algorithm to find the important features.

2. ssFCM training

   (a) Train with the selected features are used.

   (b) $\mathbf{F}$ and $\mathbf{U^0}$ are initialised based on labelled data.

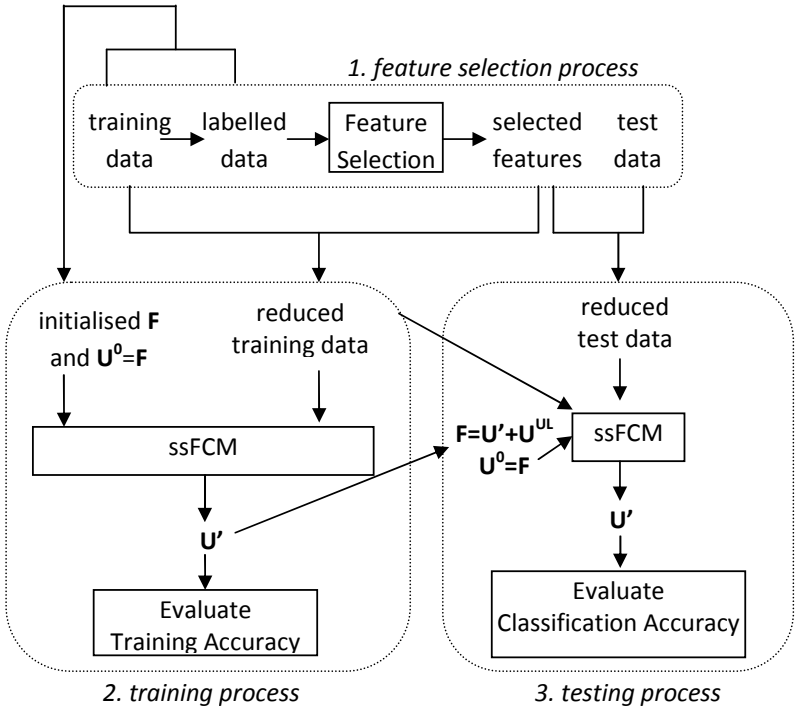


Figure 4.3: Methodology using feature selection to improve ssFCM classification

(c) The ssFCM algorithm is run. The output is the partition matrix from the last iteration, $\mathbf{U}'$.

(d) The training accuracy is then evaluated.

3. ssFCM testing/ prediction

(a) The train and test data with selected features are combined.

(b) $\mathbf{F}$ is now made up of output from training and membership matrix of unlabelled testing data $\mathbf{U^{UL}}$. The initial partition matrix $\mathbf{U^0}$=$\mathbf{F}$.

(c) ssFCM is run and the output $\mathbf{U}'$ obtained.

(d) The classification accuracy is then evaluated.

A data pattern is classified based on the highest membership value it has to a class. These class assignments are then compared with known class labels, for instance, in the case of NTBC , classifications by Soria *et al.* [141], and the number of matches are counted.

So far, important feature set based on different labelled data for each run is identified. To identify important features for the dataset, feature sets from the best performing ssFCM-feature selection methodology are chosen. In this case, the methodology should achieve highest average accuracy and stability with nf number of selected features.

Stability measures have been used as evaluation of feature selection techniques. If a feature selection technique consistently selects the same features, it builds confidence in the importance of selected features.

To calculate stability, the stability measure by Kalousis *et al.* in (2.28) is used. Next, scores based on the frequency and rank of selected features that achieved 100% are generated in (4.1) and the selected features are ranked according to these scores. The best nf number of selected features will be chosen as the important features.

Given the frequency $freq_r$ of feature f with rank r, the score for each feature is calculated as follows:

$$score_f = \sum_{r=1}^{nf} freq_r \times (nf + 1 - r) \tag{4.1}$$

In short, the methodology used in this study bases its selection process by investigating in various distance metrics (results not shown) and feature selection techniques. The feature selection techniques being investigated are CFS, NB-RFE, RF-RFE and SVM-RFE (which have been reviewed in Chapter 2.7). Furthermore, the features that give the highest average accuracy results were selected from those first selected by feature selection techniques. This means that ssFCM did not have to permute through all the combinations of 25 features to search for important features for each run. For NTBC, we experimented with 10, 15 and 17 features. One reason for choosing 10 features is that an earlier work by Soria *et al.* [140] has used this number. Furthermore, without arbitrarily specifying the number of features, different feature selection techniques will choose different number

Table 4.3: UCI dataset specifications showing number of data patterns (N), number of dimensions (n) and number of classes (c)

Dataset	N	n	c
Arrhythmia	420	278	3
Cardiotocography	2126	21	3
Yeast	1484	8	10

of features. Thus, comparisons of accuracy and stability based on the number of features will be difficult to analyse. For NTBC, a 100% accuracy result is "optimal" and means the solution completely matches with Soria's classification. Lower accuracy means less matches (similarity).

The specifications of the three UCI datasets are shown in Table 4.3. In Arrhythmia, feature 14 is removed as it contains many missing values. Data patterns in class 2 to 15 have been combined together as class 2 as there is too little data patterns in classes 7, 8 and 11 to carry out 10-fold cross validation properly. 22 data patterns which are unclassified are classed as class 3. As there is not enough data patterns from some classes to be split into 10 folds in Yeast, we carry out 2-fold cross validation. For Arrhythmia, experiments with 5, 10, 15, 20 and 25 features are carried out. For Cardiotocography, experiments with 5, 10, 15 and 20 features are carried out and for Yeast, 4 to 7 features are experimented with.

Through comparisons of average accuracies obtained by ssFCM-feature selection methodologies with ssFCM alone, we identify whether if there is an increase in average accuracy, suggesting performance improvement.

4.2.3 Results

Table 4.4 on page 99 shows the average accuracy of using feature selection with ssFCM on the NTBC dataset. To avoid clutter, we show only results from using 10, 15 and 17 selected features. Accuracy using popular filter techniques such as Info Gain (IG), Gain Ratio (GR) and Chi Square (CSQ) where the goodness of the feature subsets are evaluated based on these respective statistical measures were also compared.

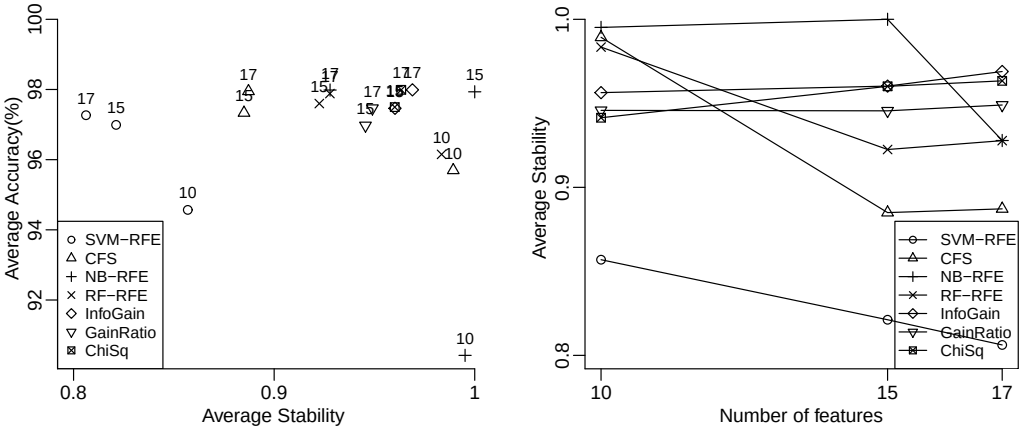
Table 4.4 on page 99 shows that there is increase in average accuracy for ssFCM using 17 features selected by various feature selection techniques than using all 25 features. NB-RFE with ssFCM was found to produce the highest accuracy for both 15 and 17 features, as shown in Table 4.4. In comparison with the other feature selection techniques, only NB-RFE with ssFCM produced a higher average accuracy using 15 features while the other techniques showed lower average accuracy than ssFCM alone. Average accuracy is higher when using 17 features for other feature selection techniques. For SVM-RFE, the average accuracy did not increase at all using 10, 15 or 17 features.

The danger of overfitting in cross-validation data [115] have been reported. To ensure that there is no overfitting in the cross-validation procedure, the comparison between the training and testing accuracy are shown in Tables 4.4 and 7.3 where no or little overfitting was found.

From Figure 4.4(a) on page 100, the 15 features selected by NB-RFE produced the most favourable results with both high accuracy and stability (with point indicated by +15 located at the most top-right). In Figure 4.4(b), NB-RFE showed the highest stability for 10 and 15 features. InfoGain, GainRatio and ChiSq have increased stability with number of selected dimensions while stability of NB-RFE, RF-RFE and CFS fluctuate with number of features. In comparison, SVM-RFE was observed to produce reduced feature sets that are less stable and of lower accuracy.

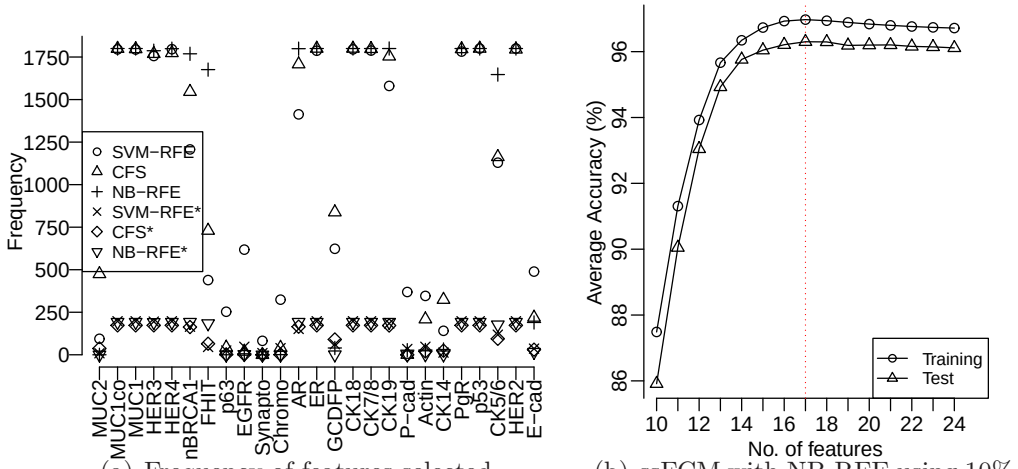
Table 4.4: Average classification accuracy of ssFCM on NTBC using all, 10, 15 and 17 selected features. Where increase in average accuracy is found, the result is indicated in italics. The highest average accuracy is indicated in bold.

	10%	20%	30%	40%	50%	60%
All features	96.12±2.04	96.86±1.94	97.22±1.77	97.54±1.61	97.64±1.55	97.84±1.53
SVM-RFE-10	90.37±4.62	92.79±3.92	94.15±3.57	94.19±3.21	94.46±3.31	94.57±3.22
CFS-10	86.34±9.15	93.59±4.44	94.68±3.14	95.10±2.89	95.38±2.91	95.70±2.46
RF-RFE-10	92.20±5.89	94.54±3.73	94.87±3.92	95.26±3.78	95.89±3.26	96.15±2.59
NB-RFE-10	85.91±7.46	87.86±7.04	88.62±5.75	89.77±4.67	89.67±5.04	90.42±4.23
IG-10	79.82±9.36	77.45±9.50	77.31±7.64	79.71±8.07	80.28±7.26	81.18±6.79
GR-10	81.60±9.85	80.53±10.32	79.72±8.66	80.13±7.80	78.69±6.15	79.30±6.27
CSQ-10	81.24±9.66	79.43±10.04	79.41±8.30	81.78±8.66	82.25±7.80	84.20±7.21
SVM-RFE-15	95.15±3.24	96.33±2.47	96.66±2.15	96.76±1.96	96.89±1.99	96.99±1.76
CFS-15	95.94±2.58	96.28±2.29	96.59±1.84	97.12±1.76	97.23±2.21	97.33±1.76
NB-RFE-15	96.05±2.47	***97.11±1.59***	***97.37±1.39***	***97.62±1.30***	***97.82±1.28***	***97.93±1.31***
RF-RFE-15	95.97±2.61	96.85±1.95	97.06±1.80	97.39±1.61	97.39±1.59	97.60±1.62
IG-15	92.55±5.81	96.30±2.40	96.73±1.80	97.12±1.60	97.24±1.49	97.46±1.58
GR-15	92.99±5.57	96.00±2.75	96.35±2.19	96.80±2.15	96.61±2.07	96.97±1.79
CSQ-15	92.60±5.81	96.18±2.69	96.73±1.66	97.16±1.64	97.28±1.54	97.50±1.48
SVM-RFE-17	95.52±2.83	96.58±2.31	96.92±1.83	97.05±1.89	97.23±1.76	97.27±1.61
CFS-17	*96.23±2.18*	*96.99±1.63*	97.07±1.52	97.40±1.57	***97.81±1.48***	*97.95±1.56*
NB-RFE-17	*96.30±2.08*	***97.14±1.73***	***97.44±1.59***	***97.70±1.47***	*97.75±1.44*	*97.98±1.37*
RF-RFE-17	*96.17±2.24*	*96.92±1.88*	*97.27±1.69*	97.53±1.58	*97.73±1.56*	*97.88±1.47*
IG-17	*96.21±2.36*	*96.97±1.55*	*97.23±1.45*	*97.60±1.41*	*97.90±1.39*	***97.99±1.36***
GR-17	***96.34±2.11***	*96.90±1.67*	97.20±1.48	97.38±1.54	97.47±1.53	97.46±1.62
CSQ-17	*96.20±2.31*	*96.93±1.60*	*97.26±1.46*	*97.56±1.40*	*97.89±1.36*	***97.99±1.43***



(a) Stability VS Accuracy (b) Stability VS Number of Features

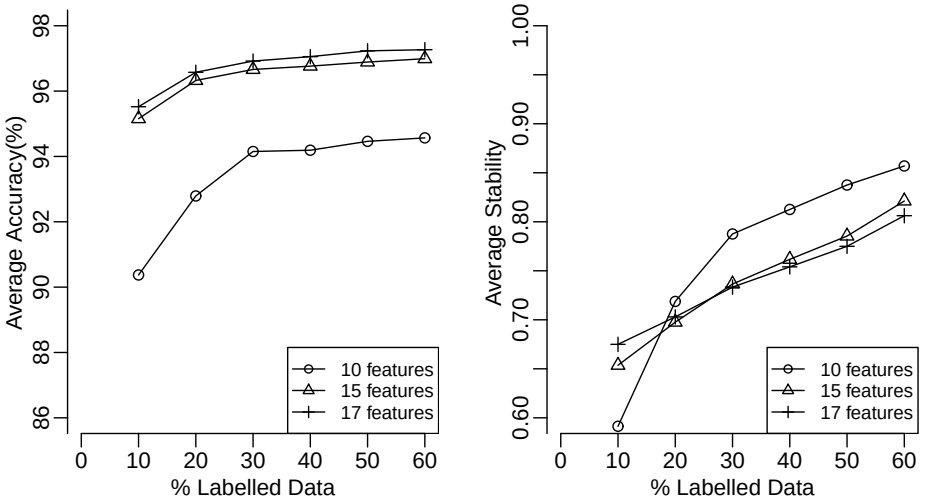
Figure 4.4: Average classification accuracy and stability of various feature selection techniques with ssFCM using 60% labelled data.



(a) Frequency of features selected (b) ssFCM with NB-RFE using 10% labelled data

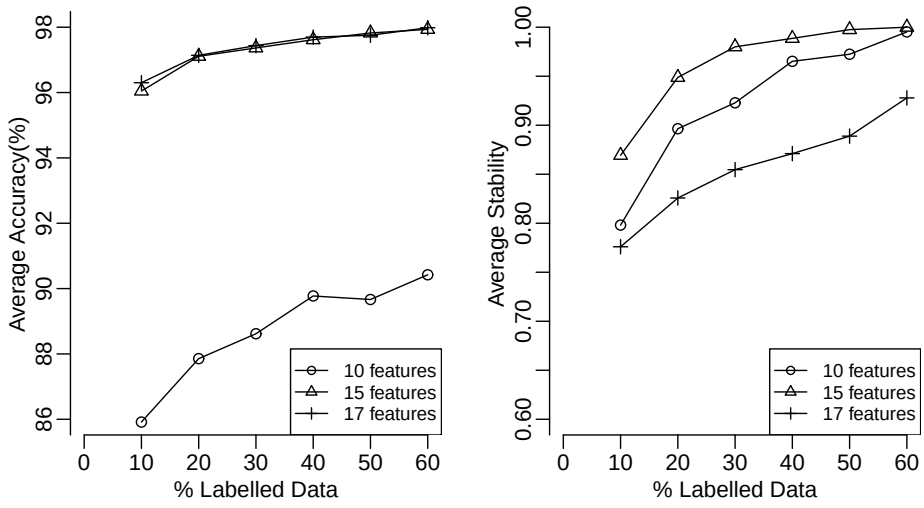
Figure 4.5: Frequency of selected features for 15 features in (a) and accuracy using 10% labelled data VS number of features in (b).

In Figure 4.5(a) on page 100, frequencies of features selected when choosing 15 features by the various feature selection techniques (only 3 techniques shown to avoid overcrowding) and the features being selected that produced accuracy of 100% denoted by * (only 2 techniques shown) are shown. ssFCM appears to disregard some of the features chosen by SVM-RFE and CFS such as MUC2 and GCDFP. While both features were occasionally selected, they did not achieve 100% accuracy with ssFCM. The lower stability found in SVM-RFE and CFS, in comparison with NB-RFE, which was reported earlier can also be observed here in greater detail where the frequency of features selected by these techniques are scattered between range of 250 and 1750.



(a) Accuracy using SVM-RFE (b) Stability using SVM-RFE

Figure 4.6: Average accuracy and stability of ssFCM with SVM-RFE on NTBC.



(a) Accuracy (b) Stability

Figure 4.7: Average accuracy and stability of ssFCM with NB-RFE on NTBC.

Figure 4.5(b) on page 100 shows the accuracy comparison between the number of features ranging from 10 to 24. The highest accuracy was achieved using 17 features selected by NB-RFE (as indicated by the red dotted line). It appears that the threshold number of features to achieve high accuracy is 14 where the knee of the graph is located and at this point, average accuracy starts to converge.

From comparison between Figures 4.6 and 4.7, for 10 features, SVM-RFE was able to produce a much higher accuracy than NB-RFE. In addi-

tion, SVM-RFE stability obtained with 10 features was higher than with 15 or 17 features when 20% or more labelled data was used. Stability is expected to increase with more features selected since the chances of selecting the same features increase as observed in Figure 4.7(b) with 10 and 15 features. But, this is not the case for using 17 features selected by NB-RFE where stability was lower than using 10 or 15 features, which suggests that different features are being selected as the 16th and 17th features. The highest stability was found using 15 features selected by NB-RFE.

To identify the important features, scores were generated based on the frequency and rank of selected features that achieved 100% with ssFCM, as shown in Table 7.4. These scores were then ranked, as shown in Table 4.5. The selected features were compared with those found by Rakha *et al.* [132]. Interestingly, EGFR, which did not appear in any of the feature lists in Table 4.5, was one of the ten important features Rakha *et al.* found. The method of selection, however, was not provided in detail.

To ensure that the feature selection methodology did not overfit, accuracy comparisons between different classifiers using 25 features and the 15 ranked selected by NB-RFE that achieved ssFCM accuracy of 100% (in Table 4.5) were conducted. The accuracy from the different classifiers are shown in Table 4.6. It showed that nine out of 11 classifiers demonstrate increased average accuracy (italicised) using the 15 ranked selected features, which is evidence that features selected by NB-RFE and ssFCM are not biased towards ssFCM alone.

Only the most favourable results (based on high average accuracy and stability) on the three UCI datasets from ssFCM with feature selection in comparison with using ssFCM alone are shown in Table 4.7 on page 105. Increased average accuracy was found using feature selection and ssFCM in all three datasets. For Arrhythmia, the most favourable results were obtained using 5 selected features by SVM-RFE. Interestingly, increase in

average accuracy was obtained using Mahalanobis distance in ssFCM with SVM-RFE, even though Mahalanobis distance with ssFCM did not produce as good results as Euclidean with ssFCM using all features. This suggests that a distance metric which previously performed worse than another distance metric using all features could give higher average accuracy when using selected features, such that it outperforms other distance metrics. For Cardiotocography, the highest average accuracy obtained in this experiment was using Euclidean ssFCM with 10 features selected by SVM-RFE. For Yeast, Euclidean ssFCM with 7 features selected by CFS produced the highest average accuracy.

Figures 4.8, 4.9 and 4.10 on pages 106 show plots of stability against accuracy and number of dimensions against stability. Note that the graphs are not in the same scale to allow clearer presentation. To avoid clutter, NB-RFE is omitted in Figure 4.8(a). The number labels on the points indicate number of dimensions. The stability against accuracy plots show which feature selection technique produce stable feature sets and which ssFCM-feature selection produce high accuracy. Ideally, the feature selection technique that produces plots at the top right corner with high stability and accuracy is chosen. This ensures that we pick a ssFCM-feature selection methodology that can produce high accuracy and that reduce set of features are consistently selected.

Figures 4.8 (b), 4.9 (b) and 4.10 (b) shows whether the stability of selected features increases with number of dimensions. Using both plots, the most suitable feature selection to be selected for use with ssFCM on a dataset can be determined.

Table 4.5: Comparison of ranked selected features from NB-RFE with those used by Rakha *et al.* in [132].

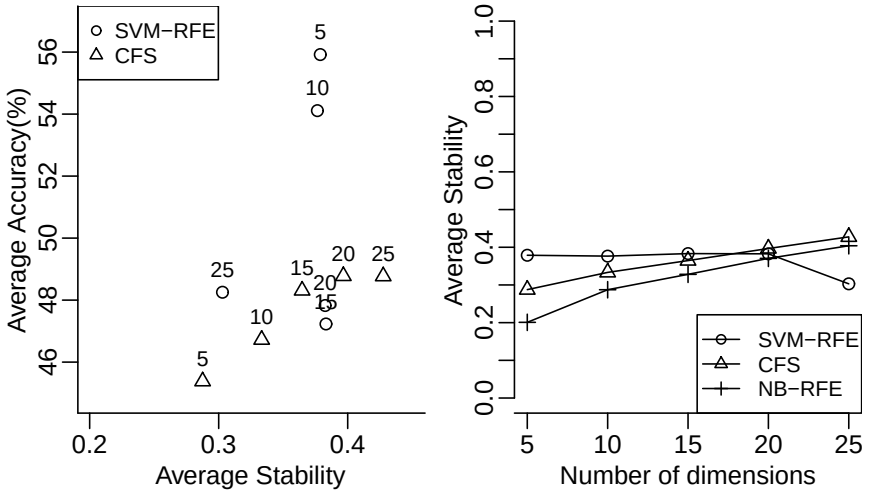
10 important features (unranked) [132]																
ER	PgR	CK7/8	CK5/6	EGFR	HER2	HER3	HER4	p53	MUC1							
rank																
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
CK7/8	CK18	p53	ER	MUC1co	MUC1	HER2	PgR	CK19	HER4							
CK18	CK7/8	p53	MUC1co	ER	MUC1	HER2	PgR	CK19	HER4	AR	HER3	nBRCA1	FHIT	CK5/6		
p53	HER2	CK18	MUC1co	ER	CK7/8	MUC1	PgR	HER4	HER3	CK19	AR	nBRCA1	CK5/6	FHIT	GCDFP	E-cad

Table 4.6: Accuracy comparison with other classifiers using feature selection. Results which shows higher average accuracy using a reduced set of features than original set are italicised.

No.	ssFCM10	GLMNET	C5.0	LDA	HDDA	KNN
25	96.12±2.04	96.50±3.07	86.42±6.06	95.17±1.57	93.97±1.72	94.58±3.02
15	*96.51±1.61*	*97.15±2.03*	85.97±5.36	*95.47±1.24*	*94.12±2.71*	*95.18±3.40*
No.	NNET	NB	MDA	KSVM	RF	LVQ
25	59.90±11.69	80.56±5.95	94.27±2.84	94.29±3.06	95.79±2.51	95.18±2.33
15	*67.02±11.45*	*92.75±2.67*	*95.34±2.11*	*97.30±1.95*	95.19±2.68	94.14±3.68

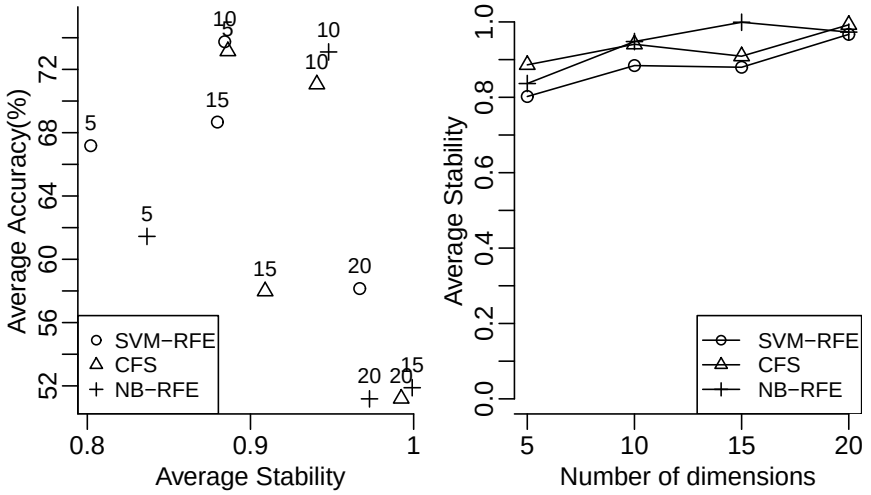
Table 4.7: Comparison of classification accuracy for the UCI datasets; Arrhythmia, Cardiotocography and Yeast, using ssFCM alone and using ssFCM with feature selection. The highest average accuracy for each dataset is indicated in bold.

Method	10%	20%	30%	40%	50%	60%
Arrhythmia						
Euclidean (E)	38.75±7.59	40.32±7.99	42.40±8.16	43.68±8.13	43.90±7.89	44.33±8.28
SVM-RFE-5+E	43.17±11.63	46.25±12.01	47.61±12.02	49.67±12.31	52.10±10.75	52.10±11.17
Mahalanobis (M)	35.37±7.42	34.13±7.43	33.25±7.39	33.62±7.03	32.68±6.48	32.84±6.72
SVM-RFE-5+M	**43.30±9.72**	**46.63±10.46**	**49.22±9.88**	**52.13±10.42**	**53.99±10.10**	**55.92±9.83**
Cardiotocography						
Euclidean (E)	47.60±3.91	48.92±3.00	49.31±3.03	49.94±3.05	50.44±3.08	51.18±3.00
SVM-RFE-10+E	**64.32±10.53**	**68.32±9.22**	**70.93±7.17**	**72.59±6.02**	**74.12±4.69**	**73.75±5.02**
Mahalanobis (M)	52.55±16.09	55.60±16.42	58.56±15.41	59.23±15.75	61.42±15.18	64.37±13.38
NB-RFE-10+M	51.51±4.62	58.00±5.15	63.11±5.24	66.78±5.07	68.97±4.61	71.18±4.09
Yeast						
Euclidean (E)	33.34±3.51	35.28±2.99	36.94±2.72	37.67±3.05	38.06±2.66	38.21±2.55
CFS-7+E	**37.70±3.25**	**39.53±2.76**	**40.64±2.48**	**41.76±2.50**	**42.09±2.67**	**43.43±2.25**
Mahalanobis (M)	34.61±2.62	35.28±2.10	36.55±2.53	37.14±2.53	37.55±3.07	37.44±2.55
CFS-7+M	37.54±1.57	38.39±1.60	39.54±2.03	40.69±1.81	41.87±1.96	42.73±2.02



(a) Stability VS Accuracy (b) Dimension VS Stability

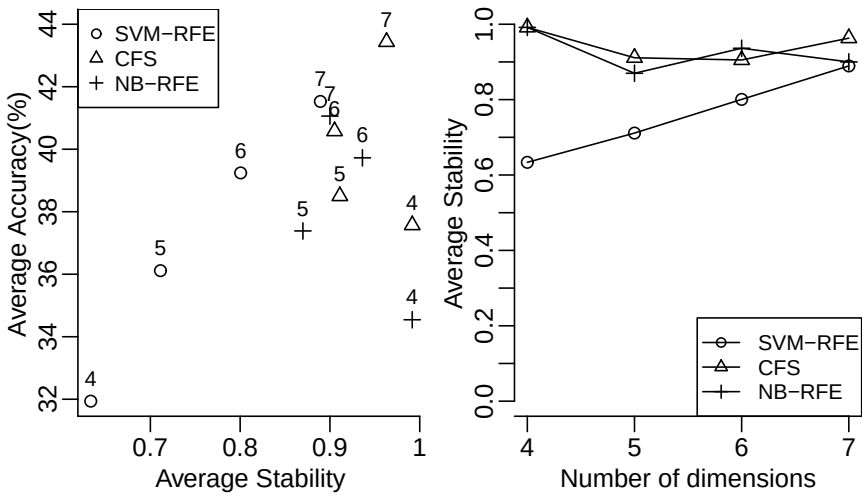
Figure 4.8: Accuracy, stability and dimension analysis for Arrhythmia dataset classification with 60% labelled data and Mahalanobis distance.



(a) Stability VS Accuracy (b) Dimension VS Stability

Figure 4.9: Accuracy, stability and dimension analysis for Cardiotocography dataset classification with 60% labelled data and Euclidean distance.

Based analysis of stability against accuracy and stability against number of dimensions plots found in Figures 4.8, 4.9 and 4.10, the most suitable feature selection techniques are SVM-RFE for Arrhythmia and Cardiotocography and CFS for Yeast. The average accuracy results are presented in Table 4.7 on page 105. Unlike NTBC where stability decreases with increasing number of features, the general trend for these datasets is that stability increases with number of features. For Arrhythmia, stability is much lower compared to the other two UCI datasets because the chances of choosing the same ones out of a high number of features are lower.



(a) Stability VS Accuracy (b) Dimension VS Stability

Figure 4.10: Accuracy, stability and dimension analysis for Yeast dataset classification with 60% labelled data and Euclidean distance.

4.2.4 Discussion

Increase in average accuracy was found using ssFCM with feature selection on NTBC and all three UCI datasets. It was observed that stability does not always increase with higher number of selected features for NTBC. Stability is related to redundancy of the feature selection technique. The low stability and fluctuations in stability found in some selection techniques may be caused by unreliable ranking of features. SVM does not account for redundancy among features [162]. SVM-RFE uses weights as the feature ranking criterion where features with small weights are retained in the selected features list even though they are redundant. Although CFS allows redundant features to be re-included in its list [71], the ranking criteria was based on goodness of the feature subset rather than on individual features. By looking at the goodness of subset, the best combination of features can be identified and their intercorrelation measured although there is some risk of redundancy. NB-RFE and RF-RFE tackle redundancy by having an outer (resampling) and inner cross-validation (wrapper approach) so that the feature subsets are tested and evaluated in the inner cross-validation

such that the subset that produces the best fit model is chosen. Like SVM, Naive Bayes cannot detect redundancy as it assumes independence between features. NB-RFE uses outer and inner cross-validation so that feature subsets are tested and evaluated in the inner cross-validation and the best fit model selected in the other cross-validation [95]. For RF-RFE, if the data contains groups of correlated features of similar importance, then smaller groups are favoured over larger groups [152].

Comparison between the 15 selected features with Rakha's ten features revealed that EGFR was not considered an important feature based on the ssFCM methodology. Apart from FHIT, the 15 ranked selected features that were generated from the ssFCM methodology were in consistent with overexpressed and underexpressed features presented by Soria *et al.* [141]. Furthermore, EGFR was reported as a modest prognostic indicator for breast cancer [116]. Further investigation to rectify EGFR's significance in breast cancer classification is required.

The danger of overfitting in feature selection techniques [67] has been reported, where the feature selection techniques may improve classification for a few classifiers. It is vital that the selected features, irrespective of the feature selection technique used, are tested on several classification techniques to evaluate the selected features. Sometimes, the selected features may be important to a particular classifier, but produce poor accuracy with other classifiers. Comparisons between accuracies by different classifiers using the features selected by the ssFCM methodology and using the original set of features showed that nine out of 11 classifiers produced higher average accuracy, indicating that the ssFCM-based feature selection methodology does not overfit.

In experiments on Arrhythmia, Euclidean ssFCM produced the highest average accuracy using all features. However, the average accuracy increased using Mahalanobis ssFCM with SVM-RFE. A similar trend was

found during tests on Cardiotocography where Mahalanobis ssFCM produced the highest results using all features, but Euclidean ssFCM with SVM-RFE produced increased average accuracy with lesser features. Based on this observation, it appears worthwhile to test the selected features on the algorithm with investigation in other distance metrics.

The increase in average accuracy found when comparing results from ssFCM with feature selection and ssFCM alone suggests that feature selection may improve classification accuracy. So far, the work done is exploratory work where a large number of techniques (distance metric, initialisation and feature selection) have been investigated in with varying amounts of labelled data, making it infeasible to conduct statistical tests on all investigations across the different datasets. Full statistical analysis will be provided in more detailed studies on a specific case study, the NTBC dataset, using different ssFCM methodologies in the following sections.

4.3 Investigating ssFCM's scaling parameter α

4.3.1 Background and motivation

The scaling parameter α maintains the balance between supervised and unsupervised learning in semi-supervised Fuzzy c-Means (ssFCM). The interest in scaling parameter α is motivated by several analysis conducted in other versions of ssFCM. In [21], Bouchachia and Pedrycz found that the number of misclassification decreases when α increases in their ssFCM algorithm with evolving membership. A higher value of α is viewed as an indicator of higher confidence on the goodness of labelled data. Furthermore, it was shown that the clusters are better separated with higher α values. In another study [23], they showed how higher α values improve results as purity (ratio of the highest number of data patterns having the same labels to the total number of data patterns in that cluster) increases

and entropy (distribution of labels in a cluster) decreases on three datasets, Diabetes, Wine and Cancer. A high purity value indicates more data patterns of the same label within each cluster and low entropy value indicate that each cluster produces a good split as it contains exclusively data patterns of one particular label in comparison with data patterns of other labels within that cluster. Bouchachia and Pedrycz [23] also showed the different effects in noise detection of the algorithm with varying values of α and using several distance metrics. However, high values of α are not always favourable. In [59], Gao and Wu presented that α values in the range [0.05, 0.2] gave the best clustering accuracy on the Iris dataset using their pairwise-constrained ssFCM. Values above 0.2 gave less favourable results. Wang *et al.* [155] expressed challenges in selecting suitable α values for pairwise-constrained ssFCM for different datasets where the range α values can be very large.

From existing studies, it is established that high values of α can have different effects on results for different ssFCMs. Thus, it is of interest to khow how the changes in α affect the results of Pedrycz97 [121] using different distance metrics for NTBC dataset and further ascertain if these effects are prevalent in other popular datasets. By knowing these effects, better selection of α values can be made to improve accuracy, taking into account the distance metric and dataset used. Furthermore, it is of interest to this research to know if α can have favourable effects on ssFCM with other methodologies such as with KKZ and feature selection (NB and ssFCM) which were previously investigated.

4.3.2 Experimental methods

In this study, the effects of different α values, 0.1, 0.5, 1, 10 and 20 in Pedrycz and Waletsky's ssFCM [121] are investigated with various amounts of labelled data, 10%, 20%, 30%, 40%, 50% and 60%. Three distance

metrics Euclidean, Mahalanobis and kernel-based are experimented with on the Nottingham Tenovus Breast Cancer dataset and five popular UCI datasets Ionosphere, Page Blocks, PID, Wine and WOBC in a clustering setting with 100 runs.

Further tests are run using 10-fold CV for the NTBC dataset with α values, 10, 20, 30, 40, 50. Tests using ssFCM methodologies with KKZ (detailed in Chapter 4.1) and the 15 features identified using NB-RFE and ssFCM (detailed in Chapter 4.2) are also conducted. Experiments using ssFCM with KKZ and the 15 features where both initialisation technique and feature selection are incorporated into a ssFCM framework are performed. The same KKZ-initialised clusters as those in Chapter 4.1 with their features reduced to the 15 features identified (in Chapter 4.2) are used. The results with the most favourable α value are compared with those by the respective ssFCM methodologies with $\alpha = N/M$.

For NTBC, a 100% accuracy result is "optimal" and means the solution completely matches with Soria's classification. Lower accuracy means less matches (similarity) with Soria's classification.

4.3.3 Results

Table 4.8 on page 112 shows the results of using various α values in ssFCM with the best performing distance metric (in terms of average accuracy). Only results of α values N/M, 0.1, 1 and 10 are shown to present the trends found. There is a general trend of increased accuracy with higher α values, particularly in PID. But, the increase in accuracy is dependent on the amount of labelled data and α value, as observed in Ionosphere, Wine and WOBC. In Page Blocks, ssFCM with $\alpha = N/M$ produced higher average accuracy than with higher values of α. In Wine and WOBC, a higher α value produced higher average accuracy with more labelled data. In Iono-

Table 4.8: Accuracy using ssFCM with different distance metric and α values on UCI datasets in a clustering setting. The highest average accuracy for each dataset is indicated in bold.

	α	10%	20%	30%	40%	50%	60%
Ionosphere	N/M	**73.82±3.71**	80.52±3.23	86.12±2.48	**89.53±1.61**	91.76±1.47	93.29±1.12
(Mahalanobis)	0.1	72.43±5.15	71.97±2.38	74.28±1.23	77.26±1.29	79.79±1.08	82.89±1.28
	1	73.51±4.06	76.71±2.78	83.42±3.02	88.48±2.25	**91.85±1.62**	**93.65±1.03**
	10	**73.82±3.71**	**81.12±3.26**	**86.75±1.96**	89.13±1.52	91.08±1.46	92.52±1.34
Page Blocks	N/M	81.01±6.40	**84.42±4.90**	**86.83±3.43**	**88.64±3.22**	90.22±2.31	**92.48±1.54**
(Fuzzy Mahalanobis)	0.1	*82.46±3.88*	83.64±1.65	83.74±1.78	83.86±1.77	83.61±1.79	84.06±1.63
	1	**83.04±4.72**	84.66±3.58	85.36±3.36	86.59±3.41	86.64±2.87	88.66±2.17
	10	81.01±6.40	84.29±5.04	86.56±3.70	88.43±3.21	**90.26±2.35**	92.44±1.78
PID	N/M	**75.17±1.28**	78.89±0.89	81.77±0.90	84.70±0.91	87.38±0.86	89.95±0.70
(Mahalanobis)	0.1	72.54±1.24	74.94±0.83	76.61±0.67	77.92±0.74	79.49±0.70	80.89±0.63
	1	75.02±1.33	78.65±0.85	81.60±0.86	84.51±0.88	87.23±0.82	89.83±0.66
	10	**75.17±1.28**	**78.93±0.90**	**81.81±0.95**	**84.91±0.85**	**87.55±0.89**	**90.18±0.67**
Wine	N/M	86.90±3.93	93.97±2.25	95.45±1.89	97.31±1.31	98.47±0.98	98.97±0.80
(Mahalanobis)	0.1	**87.07±4.04**	**94.39±2.05**	**96.22±1.78**	97.23±1.31	98.39±0.94	98.79±0.87
	1	86.13±4.14	93.54±2.27	95.51±1.82	97.30±1.32	98.48±0.94	98.93±0.81
	10	86.90±3.93	93.96±2.26	95.50±1.84	97.31±1.33	**98.53±0.98**	**98.99±0.78**
WOBC	N/M	96.64±0.25	96.75±0.31	96.98±0.37	97.39±0.36	97.81±0.33	98.25±0.39
(Kernel)	0.1	96.56±0.18	**96.85±0.23**	97.00±0.28	97.27±0.30	97.44±0.30	97.63±0.32
	1	**96.66±0.21**	96.80±0.31	**97.03±0.37**	**97.43±0.36**	97.85±0.35	98.24±0.40
	10	96.42±0.26	96.67±0.33	96.88±0.37	97.35±0.36	**97.86±0.32**	**98.29±0.39**

Table 4.9: Accuracy using Euclidean ssFCM and α values on NTBC in a CV setting. Where there is increase in average accuracy when compared to ssFCM(E) with $\alpha = N/M$ setting, the result is italicised. The highest average accuracy is in bold. p-values are calculated based on comparing between using two different alpha settings for each methodology.

α	10%	20%	30%	40%	50%	60%	100%
ssFCM (E)							
N/M	96.12±2.04	96.86±1.94	97.22±1.77	97.54±1.61	97.64±1.55	97.84±1.53	97.59±1.62
30	*96.48±2.02*	*97.55±1.71*	*97.97±1.59*	*98.35±1.36*	*98.53±1.37*	**98.58±1.35**	*98.49±1.23*
p	+	++	++	++	++	++	0.16
ssFCM with KKZ (EKKZ)							
N/M	*96.50±1.95*	*97.06±1.86*	*97.43±1.72*	*97.63±1.57*	*97.73±1.50*	*97.85±1.51*	*97.74±1.47*
30	*96.86±1.91*	*97.68±1.67*	**98.03±1.57**	**98.37±1.35**	**98.55±1.38**	**98.58±1.35**	**98.64±1.11**
p	+	++	++	++	++	++	0.12
ssFCM with 15 features (ranked by ssFCM and NB-RFE) (ENB)							
N/M	*96.51±1.61*	*97.14±1.44*	*97.35±1.35*	*97.64±1.27*	*97.81±1.28*	*97.93±1.30*	97.44±0.97
30	*96.80±1.63*	*97.73±1.38*	*97.97±1.36*	*98.18±1.29*	*98.39±1.36*	*98.43±1.22*	**98.64±1.32**
p	+	++	++	++	++	++	0.06
ssFCM with KKZ and 15 features (ranked by ssFCM and NB-RFE) (EKKZNB)							
N/M	*96.80±1.56*	*97.35±1.31*	*97.46±1.33*	*97.67±1.32*	*97.81±1.27*	*97.90±1.31*	97.59±1.06
30	**97.04±1.56**	**97.80±1.37**	*97.98±1.34*	*98.18±1.29*	*98.40±1.30*	*98.43±1.20*	**98.64±1.32**
p	0.08	++	++	++	++	++	0.10

$^{++}$ Highly significant improvement p<0.01

$^{+}$ Significant improvement p<0.05

sphere, however, $\alpha = 1$ increased average accuracy with high amount of labelled data while high α produced higher average accuracy with small amount of labelled data.

Similar trends were found in PID where accuracy increased with higher α values in ssFCM classification on the NTBC dataset, as shown in Table 7.5 on page 198. This prompted further investigation on the application of different α values in the different ssFCM methodologies previously investigated, including the extended experimentation combining initialisation technique and feature selection into ssFCM.

Table 4.9 on page 113 shows classification results using different ssFCM methodologies with $\alpha = 30$ on NTBC in a CV setting. The highest average results are in bold and average results that are higher than those produced using the original setting (in the first line of the table) are in italics. A two-sided Mann-Whitney test [131] was used to demonstrate significant improvement between using $\alpha = N/M$ and $\alpha = 30$ where the p-values (indicated in italics) are presented in the table. Significant increase in accuracy was found in all ssFCM methodologies with $\alpha = 30$ but not when amount of labelled data is 100% of training data.

A comparison between different ssFCM methodologies reveals that, with $\alpha = 30$, ssFCM with KKZ outperformed ssFCM with feature selection and produced the highest results between 30% to 60% labelled data. Previously with $\alpha = N/M$, ssFCM with feature selection outperformed ssFCM with KKZ. The combined KKZ and 15 ranked features in ssFCM with $\alpha = 30$ produced the highest results when there are 10% to 20% labelled data. At 100% labelled data, ssFCM with 15 selected features with $\alpha = N/M$ and $\alpha = 30$ produced results with lower accuracy than at 60% labelled data. Lower accuracy was also found using ssFCM with $\alpha = N/M$.

Table 4.10 show statistical significance that the ssFCM methodologies proposed perform better than Euclidean ssFCM. However, no compelling

Table 4.10: Significance test based on Mann-Whitney Test between Euclidean ssFCM and the listed ssFCM methodologies. The results show there is highly significant improvement using E30, EKKZ30, ENB30 and EKKZNB30.

	10%	20%	30%	40%	50%	60%	100%
E30[1]	+	++	++	++	++	++	0.16
EKKZ[2]	+	0.28	0.16	0.50	0.40	0.94	0.78
EKKZ30[1]	++	++	++	++	++	++	0.10
ENB[3]	+	+	0.10	0.14	+	0.26	0.96
ENB30[1]	++	++	++	++	++	++	0.12
EKKZNB[4]	++	++	++	0.06	+	0.32	0.76
EKKZNB30[1]	++	++	++	++	++	++	0.12

[1] ssFCM methodologies with $\alpha = 30$
[2] ssFCM with KKZ
[3] ssFCM with 15 features (ranked by ssFCM and NB-RFE)
[4] ssFCM with KKZ and 15 features (ranked by ssFCM and NB-RFE)
[++] Highly significant improvement $p<0.01$
[+] Significant improvement $p<0.05$

evidence was found that EKKZ performed better than Euclidean ssFCM at 20% or more labelled data. This is not to say that EKKZ performed worse or that no improvement was found, as observed previously on Table 4.2 on page 91. A similar case was found with ENB. Interestingly at 50% labelled data, significant improvement was found using all ssFCM methodologies apart from EKKZ.

4.3.4 Discussion

Increasing α value increases the influence of labelled data. It may seem obvious that increasing the influence of labelled data will improve accuracy. But this is not always the case. Setting a suitable α value, depending on the dataset and amount of labelled data, has been shown to increase average accuracy. In the experiments with the UCI datasets, three trends were observed in its classification results. First, increasing α increased accuracy as observed in PID and NTBC. Second, increasing α with increased amount of labelled data can increase average accuracy as observed with Wine and WOBC. Third, decreasing α with increased amount labelled data was observed to bring increased average accuracy for Ionosphere.

Further investigation in the effect of α with ssFCM methodologies on NTBC shows that $\alpha = 30$ improved accuracy. At $\alpha = 30$, EKKZ further improved accuracy, and outperformed all methodologies between 30% to 60% labelled data while EKKZNB30 outperformed all methodologies for lower amount of labelled data, at 10% and 20% labelled data. The higher α value of 30 further improved accuracy, accuracy that was already improved by incorporating other methodologies into ssFCM. More importantly, with $\alpha = 30$, results from ssFCM are competitive with those from EKKZ between 30% to 60% labelled data. This suggests that increasing $\alpha = 30$ can achieve accuracy as high as those achieved by incorporating other methodologies into ssFCM when 30% to 60% labelled data are available.

Therefore, for NTBC, setting $\alpha = 30$ further improved classification of ssFCM methodologies, depending on the amount of the labelled data. When amount of labelled data were between 30% to 60%, classification with ssFCM alone improved and can be as competitive as the highest results obtained from EKKZ. When labelled data were $\leqslant 20\%$, the highest results can be obtained using EKKZNB30. Based on experimental observation, a decision making strategy for using suitable ssFCM methodologies on NTBC with $\alpha = 30$ based on the availability of labelled data is illustrated in Figure 4.11. Here, when the availability of labelled data was more than 20%, the highest accuracy can be obtained using EKKZ with $\alpha = 30$.

Lower accuracy with 100% labelled data than 60% was found for some ssFCM methodologies with $\alpha = 30$, which suggests that a complete supervised setting, even with $\alpha = 30$, do not always produce favourable results. Caution has to be practised on the amount of labelled data used and the α settings. The lower accuracy in ssFCM when using $\geqslant 60\%$ of labelled data was already reported in the comparison with TSVM on page 82.

In a separate unpublished study (not included in this thesis), a combined ssFCM methodology with $\alpha = N/M$ setting was used, running KKZ

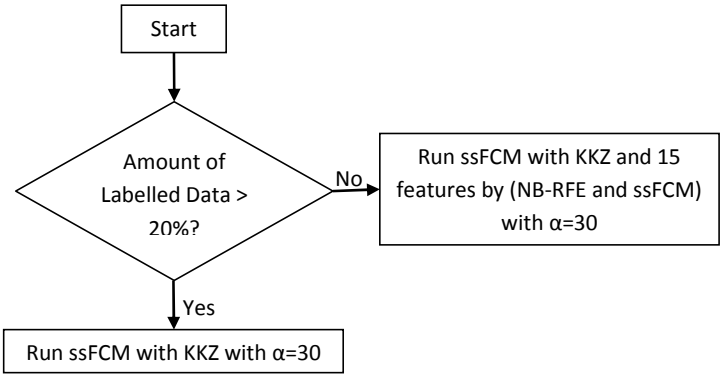


Figure 4.11: A decision making strategy for using suitable ssFCM methodologies on NTBC based on labelled data availability.

first and then, applying feature selection (and vice versa in another experiment) with ssFCM. However, better classification results are obtained from using ssFCM with KKZ applied first then feature selection.

4.4 Summary

Investigations using three initialisation techniques, SCS, KKZ and CE with ssFCM on UCI datasets and on NTBC were conducted. Increase in average accracy was found to be most apparent where the availability of labelled data were very low, indicating the use of initialisation techniques in this situation can greatly help improve classification accuracy. KKZ generates initial clusters that are at the edge of every potential cluster, allowing a better initialisation for all clusters than SCS or CE, which subsequently resulted in higher accuracy. A slight increase in accuracy was found when more than 50% of labelled data are used, which is still considered of great importance when dealing with biomedical data.

Various feature selection techniques with ssFCM were investigated in terms of accuracy, stability and overfittingness. Feature selection techniques were found to increase average accuracy in ssFCM classification on the NTBC dataset. NB-RFE with 15 features was found to produce the most favourable results in terms of both accuracy and stability. Furthermore, the best 15 features are generated based on rank and frequency of

being selected, that can achieve 100% accuracy with ssFCM. To ensure these features do not overfit or are not biased to a small number of classification techniques, these 15 features were tested with different classifiers. Nine out of 11 classifiers showed increased average accuracy. The experimental results using feature selection with ssFCM on three UCI datasets also showed that feature selection increased average accuracy in ssFCM classification for these datasets.

The effects of scaling parameter α in different ssFCM methodologies on several UCI datasets and NTBC were investigated. It was observed that setting a suitable α value, depending on the dataset and the amount of labelled data, can increase classification accuracy. Three trends in classification accuracy were identified with respect to α and amount of labelled data for the different UCI datasets. Further investigation were conducted with respect to α using 10-fold CV on NTBC, where the different ssFCM methodologies previously investigated were used. Accuracy significantly improved using all ssFCM methodologies with $\alpha = 30$ when availability of labelled data were between 10% to 60% of training data. Furthermore, the highest results can be obtained using $\alpha = 30$ in ssFCM with KKZ when amount of labelled data were between 30% to 60% and 100% and in ssFCM with KKZ and 15 ranked features when amount of labelled data were $\leqslant$ 20%. Based on these observations, a decision making strategy was built for using suitable ssFCM methodologies on NTBC based on labelled data availability.

5 From Clustering To Classification

Previously, ssFCM has been shown to classify the NTBC dataset (663 patients) with high agreement with Soria's classification using the Euclidean distance. It has been demonstrated that higher accuracy can be achieved using two approaches of improvement; initialisation and features selection techniques. Furthermore, by setting the scaling parameter α in ssFCM to 30, the accuracy improves.

This chapter demonstrates how the different investigations that have carried out previously fit together into an integrated framework. The objective of the integrated framework is to classify unlabelled or new patients such that the classification result can assist clinicians in decision making.

5.1 Background and motivation

The ultimate goal of performing classification (prediction of class labels) of unlabelled or new patients in the NTBC dataset is to provide decision making support to clinicians. As Bair and Tibshirani [9] have discussed, biological data and clinical data are often processed separately. It is thus important that the known subgroups new breast cancer patients are assigned to based on biological (immunohistochemical) data have relevance to clinical data such as survival outcome, grade and NPI. The relevance between the subgroups and clinical data not only help determines whether the subgroups patients belong to have good or poor survival outcome, but also provides a deeper understanding of the protein biomarkers that characterise the different breast cancer subgroups. Furthermore, based on the characteristics of the subgroups and their relevant survival outcome, specialised treatments for each subgroups can be administered.

In building an integrated framework based from previous investigations, the hope is to fulfill the third aim of this research, that is, to develop a framework that allows the classification of breast cancer subgroups, which can assist clinicians in providing decision making support.

To fulfill the aim of the integrated framework, it must be able to:

1. Reproduce the same six classes using the 663 classified patients data and classify new or unlabelled patient data (in this case, the remaining 413 patients) into subgroups that are biologically meaningful and

2. Show relevance between biological subgroups formed by the classification of the 413 patients to their clinical data (age, stage, size, grade, NPI and survival) and compare with those based on the already classified 663 patients for confirmation of common trends already identified by Soria *et al.* in [141] and for further insights.

5.2 Strategy

Based on the studies in Chapters 3 and 4, the different approaches are incorporated into an integrated framework for automatic classification (post-initialisation) of the breast cancer patients into the six subgroups. A conceptual diagram of the integrated framework is shown in Figure 5.1. The approaches of improvement are carried out prior to running ssFCM to either generate initial cluster centres or determine the useful features in the dataset. These parameters generated from the approaches of improvement are then incorporated into the ssFCM framework. Within the ssFCM framework, suitable distance metric for representing the structure in the dataset is chosen and the scaling parameter α is adjusted to further improve the performance of ssFCM. The ssFCM methodologies are extended to include various visual assessment and statistical techniques (or analysis) to interpret classification results in ways to show that the classes the new or

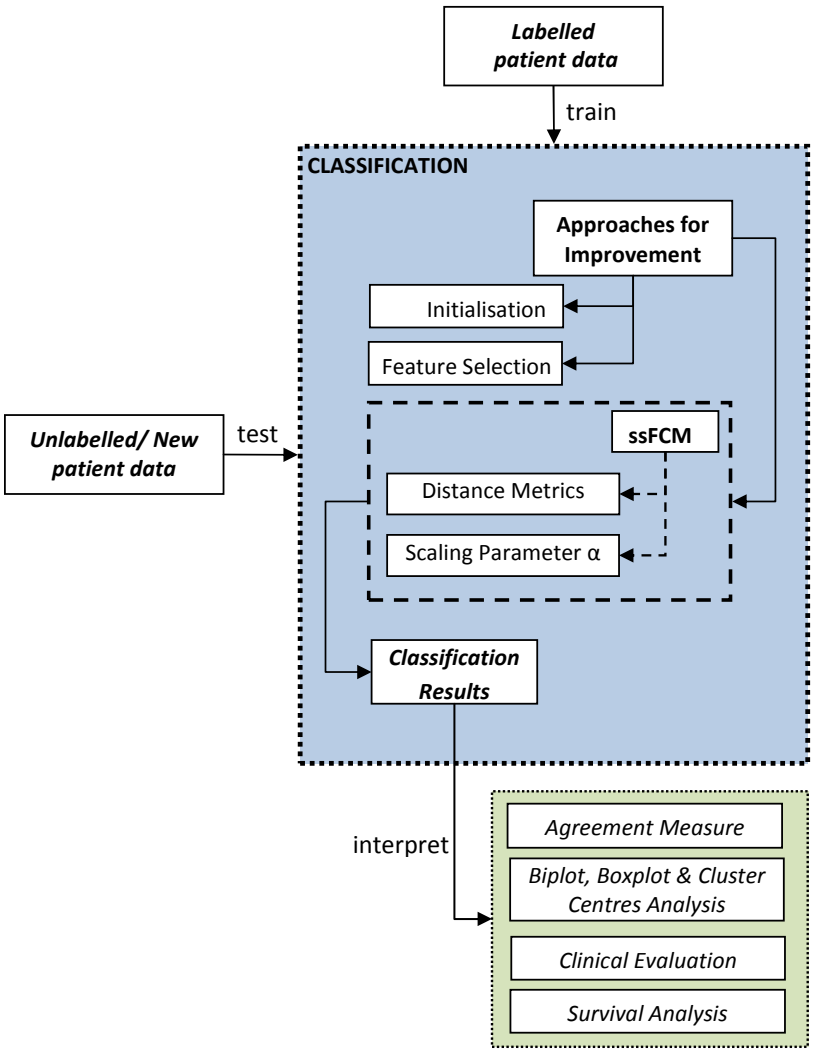


Figure 5.1: Conceptual diagram of the integrated framework.

unlabelled patients are assigned to are biologically meaningful and to show that these biological classes have relevance to clinical data that can assist clinicians in decision-making support. These approaches for interpreting classification results are illustrated in the green box in Figure 5.1.

To classify new or unlabelled data, the algorithms are trained using labelled data. The ssFCM algorithm is incorporated with two approaches of improvement. Either methodologies, ssFCM with initialisation or ssFCM with feature selection, can be used. In this case, both methodologies are ran separately and their results compared. ssFCM using both methodologies, initialisation [100] and feature selection [98], are illustrated in Figures 4.1 and 4.3 on pages 89 and 95 respectively. For ssFCM, different distance metrics were previously explored in Chapter 3 where Euclidean distance (E)

was identified to be best suited to represent the NTBC dataset. To analyse the classification results based on different α settings, both $\alpha = N/M$ and $\alpha = 30$ are used and their results compared. This is to investigate in the clinical relevance between subgroups identified by the ssFCM methodologies with $\alpha = 30$, in comparison with $\alpha = N/M$, which was originally used [121]. For classifying the 663 patients in NTBC, the highest results come from ssFCM with $\alpha = 30$ and KKZ when labelled data are more than 20%.

Initialisation techniques are incorporated to provide additional unsupervised learning to ssFCM. If the data alone (without the labels) can show where the cluster centres (prototypes) are using initialisation techniques, ssFCM can have some knowledge of what represents the clusters initially and the accuracy can be improved from having a good start. Initialisation based on labelled data may not be as accurate due to the availability and goodness of the labelled data. The available labelled data may not provide a good initial representation of the clusters. Cluster centres found by initialisation techniques can be compared with Soria's subgroups as a form of validation of Soria's subgroups. In this case, KKZ is used as it has produced the highest results with ssFCM (as demonstrated in Chapter 4.1). As KKZ is unsupervised, it is performed on the entire dataset of 1076 patients and the cluster centres are used to initialise those in ssFCM.

Feature selection is incorporated to reduce the number of biomarkers while maintaining accuracy in classifying patients. The proposed ENB methodology have improved accuracy (see Chapter 4.2). Using feature selection, 15 key protein biomarkers are identified for classifying the NTBC dataset. Furthermore, the reduction in number of protein biomarkers used for classification decreases the number of clinical tests taken to obtain data, thus, reducing time and expense and improving efficiency. Here, the same 15 important biomarkers identified using the NB-RFE and Euclidean ssFCM feature selection methodology detailed in Chapter 4.2 are used.

The six classes identified by Soria *et al.* [141] have been shown to be biologically meaningful. For this reason, Soria's classification is used as benchmark for comparison. Biplots, boxplots and cluster centres of the six classes based on their protein biomarkers distributions are analysed and compared with those by Soria *et al.* [141]. The analysis can help to determine whether the classes identified are biologically meaningful and to highlight anomalies in the classes, if present, further providing insight about the classes.

To show relevance between the biological classes (classification result) to clinical data, clinical evaluation and survival analysis are conducted. This allows the clinical characteristics and survival outcomes associated with the biological classes to be determined.

5.3 Experimental methods

Based on the strategy proposed in Figure 4.11 on page 117, ssFCM with KKZ (EKKZ) is used as there are more than 20% of dataset labelled and the objective is to achieve the highest accuracy. For comparison, the experiment is conducted twice, one with N/M setting of α and another with $\alpha = 30$. For further investigation in the difference in results, the experiment using ssFCM with the 15 features selected by NB-RFE and ssFCM (ENB) is conducted. All methodologies were able to retain the whole labelled data completely, meaning 100% training accuracy. The ENB methodology used to find the 15 important features is detailed in Chapter 4.2.

The class labels assigned to these data patterns are based on the highest membership value it has to a class. The results for the 413 patients are referred as EKKZ or ENB classification based on their respective methodologies. EKKZ30 and ENB30 refers to classification with $\alpha = 30$.

The classification obtained from the two methodologies with different α settings are compared using confusion matrix and agreement measure.

Table 5.1: Interpretation of Cramer's V association measure based on [38].

Cramer's V value	Interpretation
0 to 0.1	Very weak if any association
0.1 to 0.3	Weak association
0.3 to 0.5	Moderate association
>0.5	High association

There are no prior labels to evaluate the correctness of the classification, nor to evaluate the performance of the framework. Thus, visual comparisons of boxplots, biplots of biomarker distributions, and cluster centres across the 6 classes with those by Soria and colleagues [141] are conducted for evaluation. The cluster centres for the six classes are determined by calculating the average of biomarker values for each class based on the classification results of the 413 patients.

To demonstrate that the framework is capable of providing decision-making support, clinical evaluation is performed by investigating in the correlation between the class distributions and clinical parameters (such as age and stage) using Cramer V. The Cramer V [37] is an association measure between two nominal variables, where both variables can have more than 2 classes. It measure ranges from 0 to 1 with 0 indicating no association and 1 for complete association. The interpretation of Cramer's V values are presented in Table 5.1. This is implemented using the `assocstats` function in the `vcd` R package [112].

Associations between the classifications with survival are analysed using Kaplan-Meier analysis. Surviving patients with less than 60 surviving months are not used in the survival analysis as their outcome after the 60 months are currently unknown. Those that have unknown survival status are also not included in the analysis. These conditions are enforced to ensure a more realistic survival analysis. The class distribution of patients used in the survival analysis are tabulated in Table 7.6 in the appendix on page 199. There are three main groups of breast cancer which are further

Table 5.2: Confusion matrix of classifying 413 patients using EKKZ (rows) and ENB (columns). The table shows high number of matching classification.

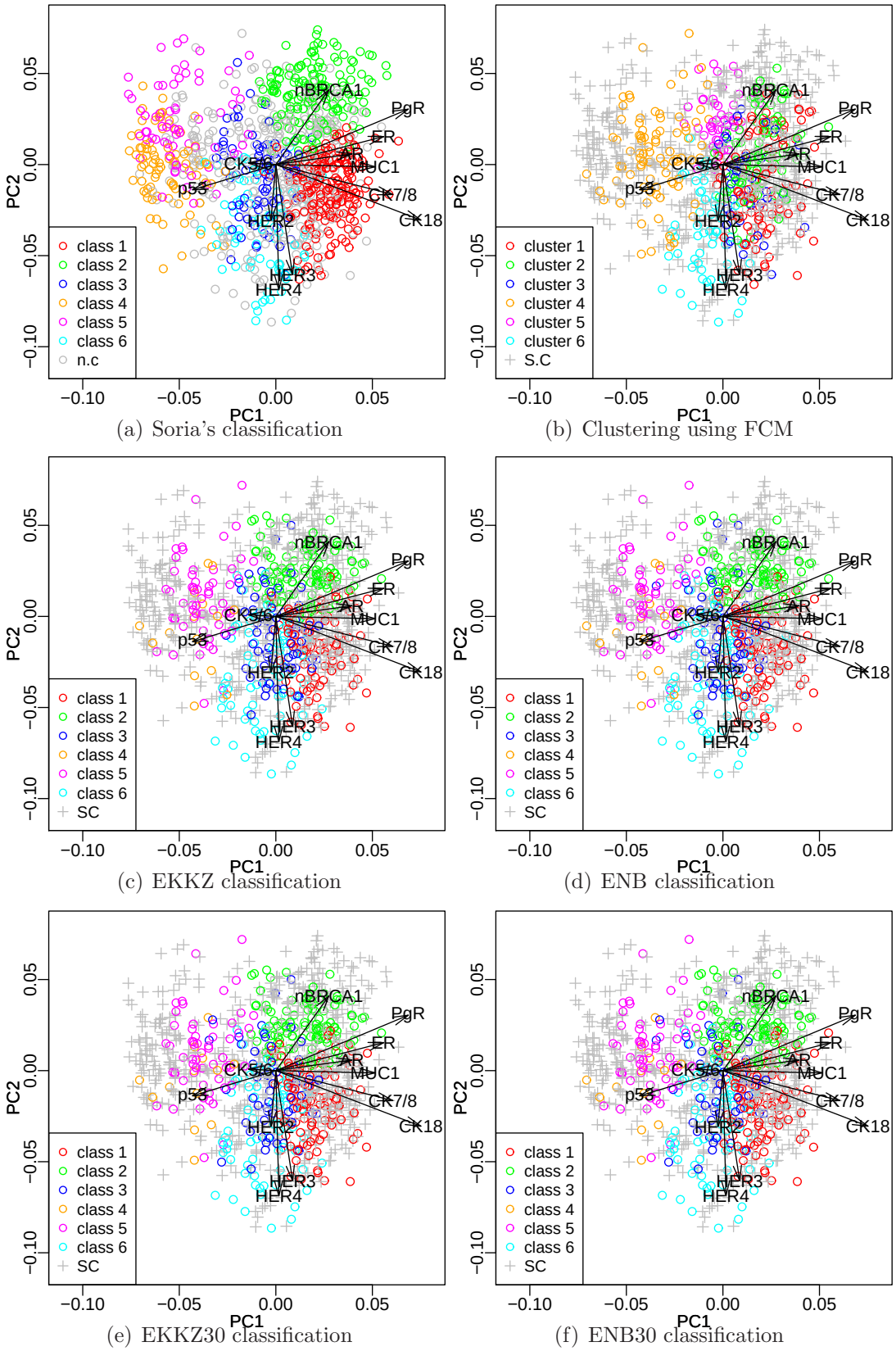
	ENB						
EKKZ	1	2	3	4	5	6	total
1	85	5	0	0	0	1	91
2	8	102	1	0	0	2	113
3	1	1	62	0	1	1	66
4	0	0	0	15	0	0	15
5	0	0	1	0	57	0	58
6	1	0	0	0	0	69	70
total	95	108	64	15	58	73	413

subdivided into six subgroups [3, 141]. Furthermore, it has been shown that the Basal subgroups – classes 4 and 5 and HER2 subgroup – class 6 are associated with poor prognosis. If the survival curves for the six classes can demonstrate similar survival trends as in [141], it can ascertain the framework to be capable of providing decision-making support.

NPI boxplots are drawn up and compared with those by Soria *et al.* [141] to study association between the biological classes and NPI. High NPI values are often associated with poor prognosis. If classes 4-6 are shown to have higher NPI values, this is another demonstration of the framework's capability as a decision-making support tool.

5.4 Results

Table 5.2 on page 125 shows the confusion matrix of classifying the 413 unlabelled patients using EKKZ and ENB. They both show highly similar classification solutions achieving a κ agreement of 0.931. There is some dissimilarity in classes 1, 2 and 6. Table 5.3 shows the confusion matrix of classifying the 413 unlabelled patients using EKKZ30 and ENB30. Interestingly, they show near complete agreement with Cohen's κ agreement of 0.995. With $\alpha = 30$, the classification results from both techniques are almost the same, indicating more stable solutions. EKKZ30 and ENB30 completely agreed on classes 4-6 assignments of patients.



(a) Soria's classification (b) Clustering using FCM

(c) EKKZ classification (d) ENB classification

(e) EKKZ30 classification (f) ENB30 classification

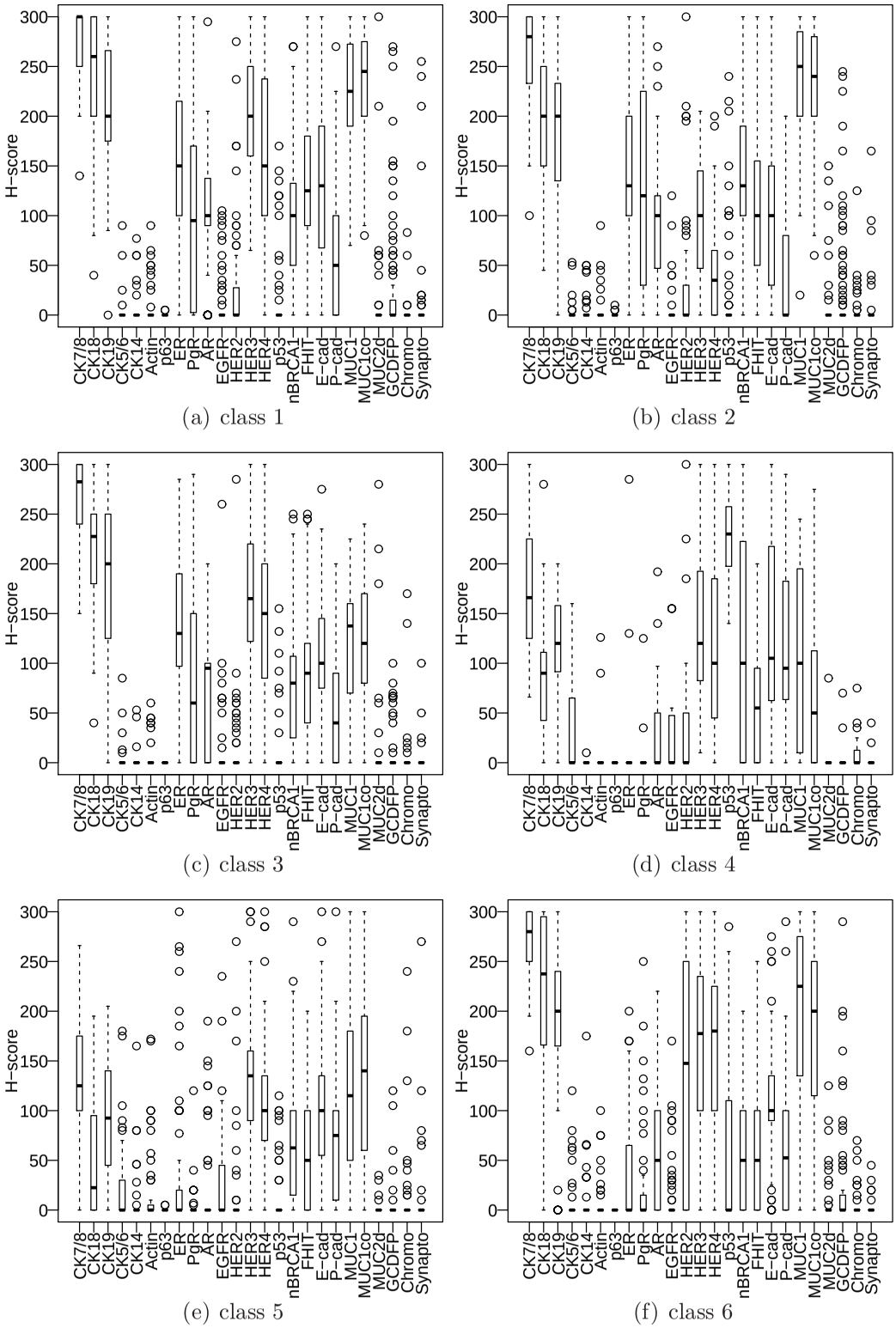
Figure 5.2: Biplots showing Soria's classification (SC) [141] and not classified (n.c) patients in (a), clustering of 413 patients into 6 clusters using FCM in (b), the classification of 413 patients using EKKZ in (c), and using ENB in (d), using EKKZ30 in (e) and ENB30 in (f).

Table 5.3: Confusion matrix of classifying 413 patients using EKKZ30 (rows) and ENB30 (columns). The table shows high number of matching classification.

	ENB30						
EKKZ30	1	2	3	4	5	6	total
1	107	0	0	0	0	0	107
2	1	110	0	0	0	0	111
3	1	0	55	0	0	0	56
4	0	0	0	13	0	0	13
5	0	0	0	0	59	0	59
6	0	0	0	0	0	67	67
total	109	110	55	13	59	67	413

Figure 5.2 on page 126 shows biplots of Soria's classification in (a), clustering result using FCM in (b), classification of 413 patients using EKKZ in (c), using ENB in (d), EKKZ30 in (e) and ENB30 in (f). The biplots were plotted using the first and second components from Principal Component Analysis (PCA). Note that PCA was used to generate 2 dimensions for visual analysis and no feature reduction was performed on the classification. Classification results from both ssFCM methodologies were shown to resemble Soria's classification. For FCM clustering result, the clusters were manually aligned with Soria's classification to enable comparison. Figure 5.2(b) shows that FCM was not able to distinguish the two clusters in the Basal region and instead clustered the whole region as cluster 4.

The boxplots in Figure 5.3 on page 128 show the distributions of the 25 protein biomarkers for the six classes based on EKKZ results. There appears to be no visible difference between boxplots of EKKZ, ENB, EKKZ30 and ENB30. Thus, only boxplots of EKKZ are shown. ENB and ENB30 were able to produce similar boxplots as those from EKKZ. For Basal subgroups in classes 4 and 5, it was observed that the CK14 expression (one of the protein biomarkers that characterises them) was low, which was contrary to that of Soria and colleague's findings in [141]. However, the boxplots for classes 4 and 5 still reflected other triple negative breast cancer characteristics with high expression of Basal CK5/6 and low expression in



(a) class 1 (b) class 2

(c) class 3 (d) class 4

(e) class 5 (f) class 6

Figure 5.3: Boxplots showing statistical summaries of all biomarkers for the six classes obtained from EKKZ. The boxplots show visual similarity to Soria *et al.* [141].

Table 5.4: Cluster centres for each class (c) using EKKZ and ENB (in brackets) with range (R) and standard deviation (SD). Where there is high value which may indicate overexpression, the result is underlined.

c	CK7/8	CK18	CK19	CK5/6	ER	PgR	AR	HER2
1	285(284)	251(249)	222(221)	1(1)	152(152)	156(152)	103(102)	16(15)
2	267(268)	198(199)	198(200)	2(2)	135(135)	164(168)	97(98)	15(15)
3	267(267)	212(213)	188(187)	5(5)	144(147)	75(74)	76(76)	21(21)
4	134(134)	45(45)	86(86)	43(43)	7(7)	2(2)	8(8)	11(11)
5	115(115)	36 (36)	77(77)	39(39)	29(28)	11(13)	15(15)	11(11)
6	258(257)	206(206)	190(190)	8(8)	27(29)	20(22)	53(53)	174(173)
R	170(169)	215(213)	145(144)	42(41)	145(145)	162(165)	95(94)	164(162)
SD	76(72)	93(87)	62(57)	19(20)	68(67)	73(75)	41(39)	65(75)

c	HER3	HER4	p53	nBR1	MUC1	MUC1c	FHIT
1	202(202)	166(166)	12(11)	96(96)	219(220)	230(230)	129(128)
2	74(73)	31(30)	10(10)	160(162)	217(215)	227(227)	105(108)
3	159(158)	137(134)	17(17)	84(84)	76(74)	80(78)	87(85)
4	157(157)	117(117)	239(239)	61(61)	91(91)	84(84)	57(57)
5	117(116)	89(89)	22(22)	76(75)	106(107)	100(101)	47(47)
6	164(162)	162(160)	81(81)	63(63)	216(215)	203(204)	67(66)
R	128(129)	135(136)	229(229)	99(101)	143(147)	149(152)	82(80)
SD	44(41)	51(47)	90(103)	37(35)	70(64)	73(67)	12(28)

ER, similar to those by Soria *et al.* [141]. For the other classes, the boxplots were found to maintain similar trends as Soria *et al.* [141] for the key features of each classes described in Chapter 2.8.2. Another observation is that the biomarker characteristics appears less distinctive as compared to those in the boxplots of Soria *et al.* [141], which explains the very reason they belong to mixed classes in the first place. For instance, class 6 (HER2 subgroup) in [141] showed boxplot of HER2 with median of above 200 and an interquartile range of above 150 and below 300. In the results here, however, the HER2 distribution of class 6 shows a median of 150 with a more dispersed interquartile range of 0 and 250. From Figure 5.2(a), the 413 patients (in grey points) were found bordering at the edges of clusters and located further away from the cluster centres. This explains their weaker characteristics (indicated by lower expression levels) than the 663 patients observed in the biomarker distributions of the six classes.

Table 5.4 shows the cluster centres which are made up of average biomarkers values that characterise the six classes, based on the classification of the

413 patients. Biomarkers with both standard deviation (SD) and range (R) of less than 40 were removed as they did not appear to help discriminate between classes. The underlined values show high expression of the biomarkers. The results by ENB are presented within brackets and values not within brackets are results obtained from EKKZ. Both techniques, as well as with $\alpha = 30$, produced very similar cluster centres for the six classes, despite slight dissimilarity in classification results as observed in Tables 5.2 and 5.3. Very high p53 and HER2 expressions for classes 4 and 6 were respectively found. Classes 1-3 have high ER values while classes 4-6 have otherwise. Class 3 showed lower PgR, which differentiated it from class 1 and 2 and class 2 had much lower HER3 and HER4, which differentiated it from class 1 and 3. These characteristics mirror Soria's classification. Due to its low range and standard deviation value, CK14 was dropped. This is no surprise as the boxplots showed low CK14 expressions for the Basal subgroups. Interestingly, these more discriminative biomarkers from EKKZ are the same 15 features selected using ssFCM and NB-RFE (see Chapter 4.2). This indicates that the cluster centres can help identify discriminative biomarkers from simple statistics, such as the standard deviation and range. However, FHIT values between the classes did not appear to help discriminate between classes.

Table 5.5 shows the class distribution of EKKZ (presented without brackets) and ENB (presented in brackets) classifications based on clinical parameters and their associations with clinical parameters measured using the Cramer's V coefficient. The Cramer's V values are presented in italics. Based on the interpretation of Cramer's V values in Table 5.1, weak association was found between the classifications and all the clinical parameters except grade. Moderate association was found between the classifications and grade. Once again, ENB was able to produce similar distributions and Cramer's V values as EKKZ.

Table 5.5: Class distribution based on clinical parameters for 413 patients. The values in italics denote the Cramer's V value with those in brackets for classification based on ENB and without brackets for classification based on EKKZ.

Parameters *Cramer's V*	cla.1	cla.2	cla.3	cla.4	cla.5	cla.6
Age 0.15 (0.14)						
⩽35	1 (0)	3 (3)	6 (6)	1 (1)	3 (3)	3 (4)
35<Age⩽45	56 (61)	62 (57)	33 (33)	6 (6)	20 (20)	32 (32)
45<Age⩽55	14 (11)	14 (16)	11 (10)	3 (3)	15 (16)	14 (15)
>55	20 (23)	34 (32)	16 (15)	5 (5)	20 (19)	21 (22)
Total	91 (95)	113 (108)	66 (64)	15 (15)	58 (58)	70 (73)
Grade 0.39 (0.38)						
1	16 (18)	29 (27)	7 (6)	0 (0)	1 1()	1 (2)
2	48 (50)	53 (51)	16 (15)	1 (1)	4 (4)	13 (14)
3	27 (27)	31 (30)	43 (43)	14 (14)	53 (53)	56 (57)
Total	91 (95)	113 (108)	66 (64)	15 (15)	58 (58)	70 (73)
Size 0.13 (0.13)						
⩽1.5cm	40 (39)	39 (39)	14 (14)	5 (5)	18 (18)	17 (18)
1.5cm<Size⩽2cm	7 (7)	13 (13)	10 (10)	0 (0)	8 (7)	15 (16)
2cm<Size⩽2.5cm	21 (22)	30 (29)	25 (25)	3 (3)	17 (17)	16 (16)
2.5cm<Size⩽3cm	11 (15)	19 (16)	11 (9)	3 (3)	11 (12)	14 (14)
<3cm	12 (12)	12 (11)	6 (6)	4 (4)	4 (4)	8 (9)
Total	91 (95)	113 (108)	66 (64)	15 (15)	58 (58)	70 (73)
Stage 0.18 (0.16)						
1	52 (56)	74 (66)	28 (30)	7 (7)	42 (41)	34 (37)
2	36 (37)	27 (29)	32 (28)	5 (5)	14 (15)	25 (25)
3	3 (2)	12 (13)	6 (6)	3 (3)	2 (2)	10 (10)
Total	91 (95)	113 (108)	66 (64)	15 (15)	58 (58)	69 (72)
Death 0.22 (0.20)						
No	87 (89)	104 (101)	60 (57)	13 (13)	49 (50)	53 (56)
Yes	3 (5)	8 (6)	3 (4)	1 (1)	6 (5)	15 (15)
Total	90 (94)	112 (107)	63 (61)	14 (14)	55 (55)	68 (71)
NPI 0.21 (0.20)						
⩽ 2.4 (EPG)	12 (14)	13 (11)	3 (3)	0 (0)	1 (1)	2 (2)
2.4<NPI⩽3.4 (GPG)	25 (26)	37 (36)	14 (13)	1 (1)	3 (3)	7 (8)
3.4<NPI⩽4.4 (MPG1)	21 (22)	34 (31)	11 (13)	3 (3)	23 (22)	16 (17)
4.4<NPI⩽5.4 (MPG2)	25 (23)	16 (17)	18 (16)	6 (6)	22 (23)	22 (24)
<5.4(PPG)	8 (10)	13 (13)	20 (19)	5 (5)	9 (9)	23 (22)
Total	91 (95)	113 (108)	66 (64)	15 (15)	58 (58)	70 (73)

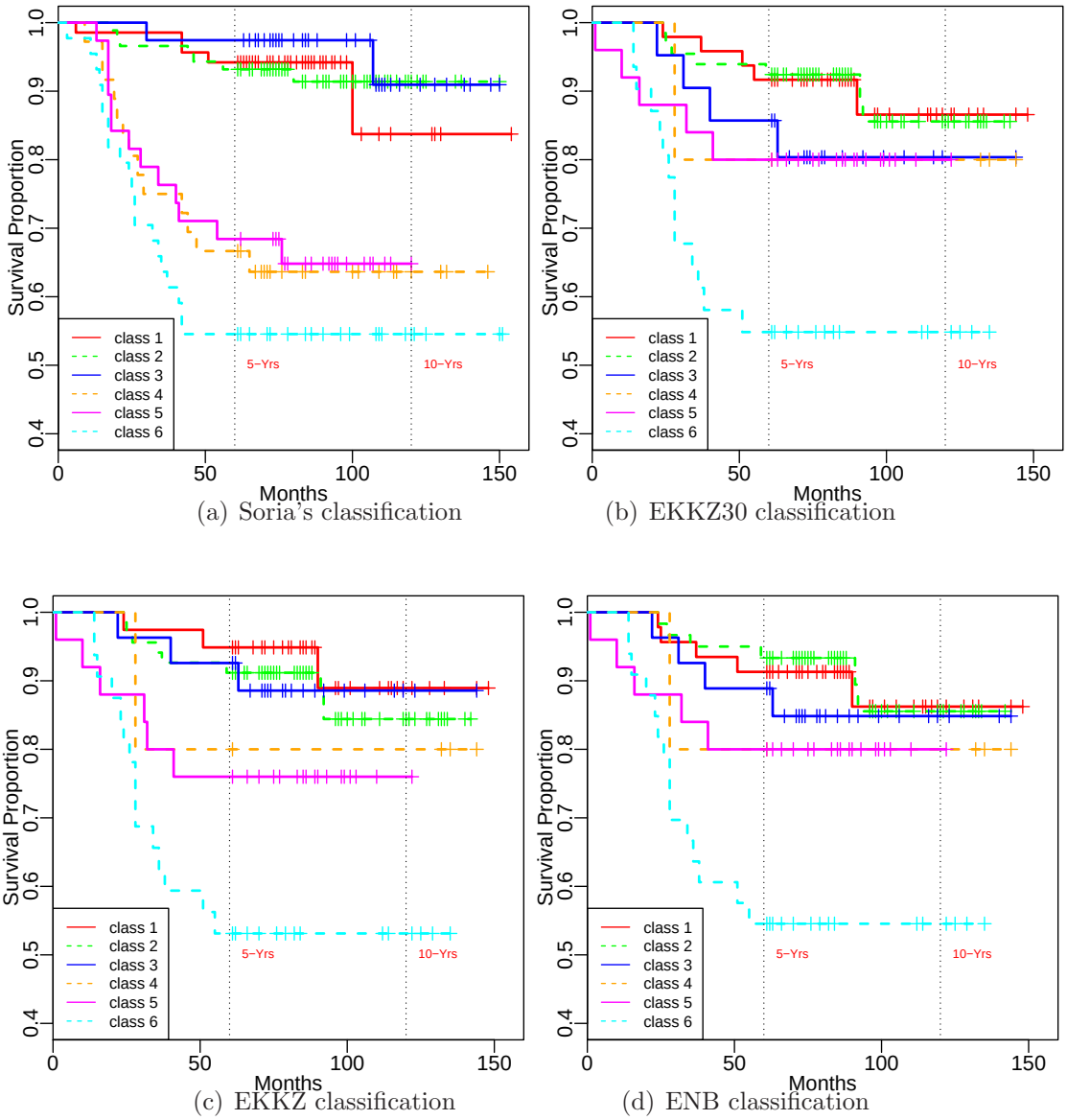
Table 5.6: Clinical evaluation by association between clinical parameters and classification based on Soria *et al.* [141], EKKZ, ENB, EKKZ30, ENB30 measured by Cramer's V.

Parameters	Soria *et al.*[141]	EKKZ	ENB	EKKZ30	ENB30
Age	0.15	0.15	0.14	0.13	0.13
Grade	0.47	0.39	0.38	0.39	0.38
Size	0.15	0.13	0.13	0.12	0.13
Stage	0.15	0.18	0.16	0.16	0.16
NPI	0.26	0.21	0.20	0.21	0.21
Death	0.30	0.22	0.20	0.20	0.20

A high percentage of patients (93%, 91% and 80%) with Grade 3 (characterised by cancer cells that grow more quickly than Grades 1 and 2) were found in the more aggressive subgroups, classes 4-6 respectively based on EKKZ classification. Also based on EKKZ classification, there was a higher percentage of deaths (22%) found in class 6 than the other classes ($\leqslant 11\%$). Similar trends were also found using ENB, EKKZ30 and ENB30.

Associations between clinical parameters with classification based on Soria *et al.* [141], EKKZ, ENB, EKKZ30 and ENB30 were compared and are shown in Table 5.6. EKKZ classification appeared to have slightly higher association coefficients (for age, stage and death) as compared to the other ssFCM methodologies. The categorisation of these clinical parameters were the same as those in Table 5.5.

The survival curves for each respective classes based on Soria's classification and classification by EKKZ30, EKKZ, ENB are found in Figure 5.4. The survival curves of EKKZ30 and ENB30 are visually similar, thus, the latter is not shown. Like survival curves of Soria's classification, based on visual analysis, survival curves of EKKZ classification at the 5-year survival time showed distinction between the three main classes and their subclasses, indicating strong association between survival outcomes and classes. However, based on the log-rank test in Table 7.8 in the appendix on page 200, the survival differences for the 6 subgroups based on pairwise comparison between each survival curve are not always significantly differ-



(a) Soria's classification (b) EKKZ30 classification

(c) EKKZ classification (d) ENB classification

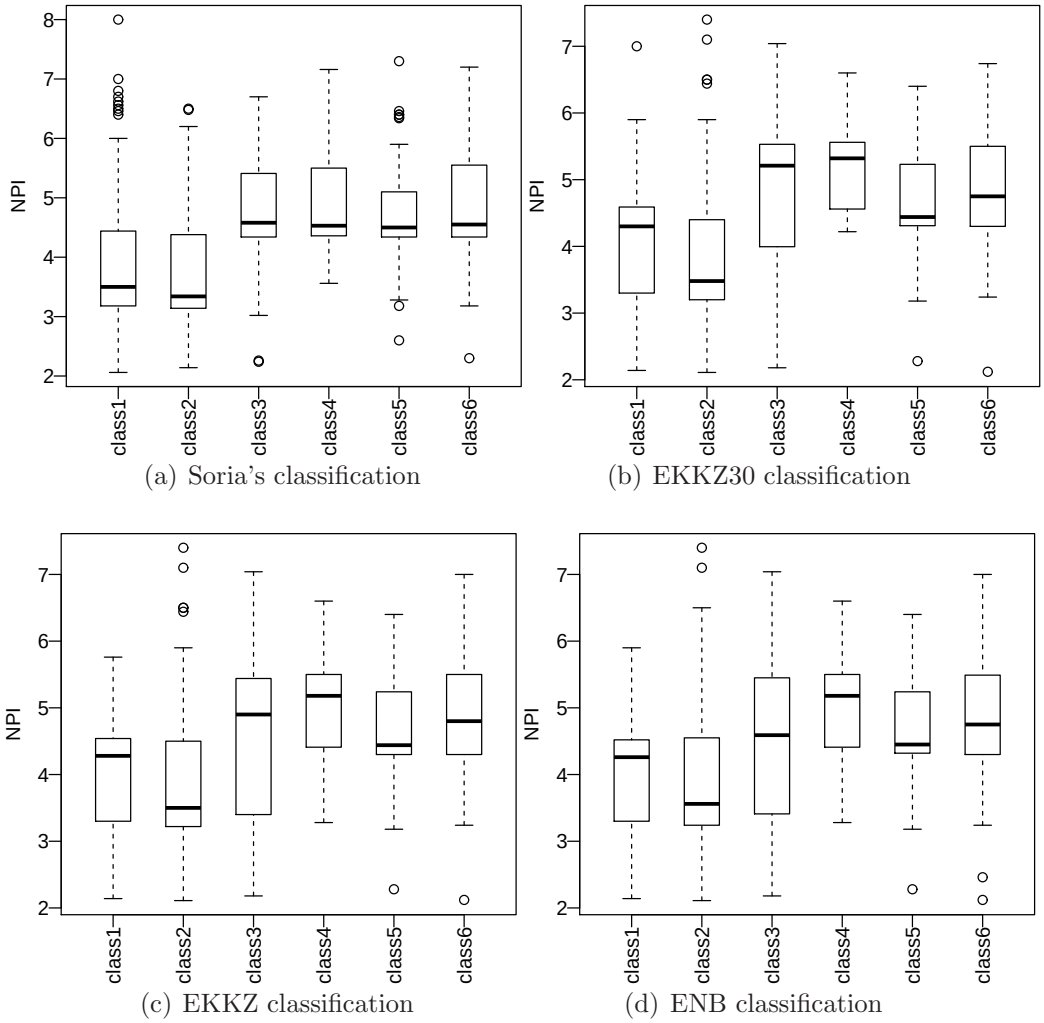
Figure 5.4: Kaplan-Meier analysis of overall survival.

ent even though overall comparison shows significant difference. There is a stronger difference between survival curves based on the 3 main groups, as indicated by the smaller p-values. The demonstration of these distinctions in patients that were previously found in mixed classes is a very positive indicator of the validity of Soria's subgroups. For ENB, the survival curves show clearly the distinction between the three main groups, but the distinction between the Basal subgroups are not clear (see Table 7.10 in the appendix on page 200 for p-values based on log-rank tests). For EKKZ30 and ENB30 in Figure 5.4(b), class 3 shows poorer survival outlook as compared to Soria's classification, EKKZ and ENB classification.

The distinction between Luminal and Basal groups are less clear and so is the distinction between Basal subgroups. Table 7.7 to Table 7.11 in the appendix on pages 200 and 201 show the difference between the survival curves based on 6 subgroups and 3 main groups for Soria's classification and classifications by EKKZ, EKKZ30, ENB and ENB30. In general, there is significant difference between Luminal (class 1) and HER2 (class 3), and Basal (class 2) and HER2 (class 3) found in the survival curves based on 3 the main groups, which indicates there is some association between the survival curves with the groups. Figure 7.1 in the appendix on page 199 shows the survival curves based on these three main groups generated by EKKZ, EKKZ30, ENB and ENB30. The survival curves show visibly clear distinction between the three groups for all four ssFCM methodologies.

For EKKZ and ENB classification, similar survival trends with Soria's were found in classes 1-3 and 6 (see Table 7.8 and 7.10 respectively for more details). Although the survival curves based on the 6 subgroups are not significantly different based on the log-rank tests, visually for EKKZ, it can be observed that there is separability between the three main classes, which is positive results, especially as the 413 patients are difficult to classify. Survival curves for classes 4 and 5 from the classification show a more optimistic survival outcome than Soria's. The small class 4 population may have produced a more optimistic outcome as there are only five class 4 patients and 25 class 5 patients could fulfill the condition to be used for this analysis for the four ssFCM methodologies.

Associations between the classifications with NPI were analysed using boxplots in Figure 5.5. The NPI boxplots show resemblance with Soria's classification for classes 1-2 and 4-6, but class 3 show a higher dispersion, as further shown using the Kruskal-Wallis test in Table 7.12 in the appendix on page 201. The test is used to determine if the NPI distribution of each class from Soria's classification is identical to those from the dif-



(a) Soria's classification (b) EKKZ30 classification

(c) EKKZ classification (d) ENB classification

Figure 5.5: Boxplots showing NPI distribution for six subgroups based on classification of 413 patients.

ferent ssFCM methodologies. Class 3 is located beside all other 5 classes (see biplots in Figure 5.2 on page 126) which may be the reason for the higher dispersion. High NPI values indicate poor prognosis, which can be observed with classes 4-6. Despite some discrepancies, these findings not only support Soria's classification, but confirm that the framework can produce an accurate prognosis. Interestingly for NPI distribution based on Soria's classification [141], NPI distribution for class 3 is higher than classes 1 and 2. However, the survival curve for class 3 shows the best survival outlook which is somewhat contradictory as high NPI value are associated with very poor prognosis [57]. For ssFCM methodologies, the high NPI distributions for classes 3-6 correspond with the lower survival curves in Figure 5.4.

5.5 Discussion

Based on observation in Table 4.9 on page 113, accuracy slightly improved using EKKZ30 and ENB30 with 100% labelled data than using EKKZ and ENB. The contrary was observed in the classification of 413 patients where EKKZ30 and ENB30 produced less clinically relevant classification than EKKZ and ENB classification, based on the separability of the survival curves which reflected their subgrouping. In this case, ssFCM with $\alpha = 30$ in a completely supervised setting (with 100% labelled data) produced less favourable results in comparison. With 100% labelled data and $\alpha = 30$, EKKZ30 and ENB30 modelled Soria's classification very similarly, which may suggest that the influence from KKZ or from feature selection is overshadowed by influence from the labelled data. Furthermore, the 413 patients were mostly located at the borders of the six subgroups (see Figure 5.2(a) on page 126). The high contribution from supervised learning, with 100% labelled data and $\alpha = 30$, may not be suitable in this case as it imposes strict cluster borders based on influence from labelled data. For these reasons, a relaxed ssFCM with $\alpha = N/M$ allows more influence in unsupervised learning of the unlabelled data and may be more appropriate for this task, as observed in Figure 5.4(c) and (d) on page 133.

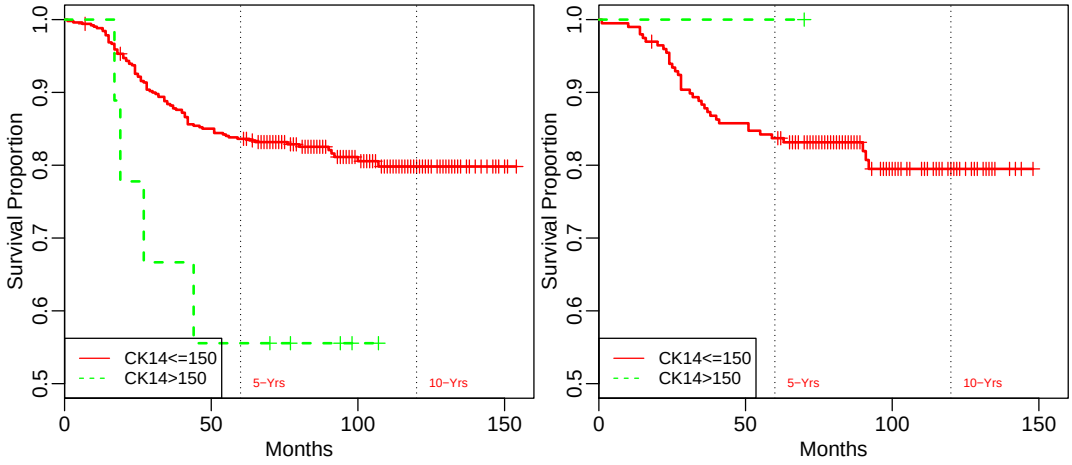
Using EKKZ and ENB, the same six classes of breast cancer types as Soria's classification can be identified for the 413 patients. This evaluation showed that the ssFCM framework was able to accurately classify the NTBC dataset. Furthermore, this confirmed Soria et al's six classes and addressed the issue of stability of their classification.

While all the features in NTBC are highly non-normally distributed, the framework has been able to produce very good classification results. Furthermore, Hair *et al.* stated that the requirement of normal distribution [69] has little effect in clustering techniques. More importantly, EKKZ

and ENB were able to detect relevant areas of high concentration (see biplots in Figure 5.2 (b) and (c) on page 126) that were of importance using 663 labelled data as training examples, irrespective of the distribution. ENB was able to achieve similar results using only 15 selected features.

The initial reasoning for the more optimistic survival outcomes in Basal groups classes 4 and 5 was due to very low expression of CK14 as observed in the biplots in Figure 5.3 on page 126, shown only as outliers. On further analysis of CK14 expressions with survival outcomes in both the 413 unlabelled and 1076 total patient groups in Figure 5.6, the reasoning that Basal subgroups with low CK expression would give a more optimistic survival outcome did not hold. Instead, the optimistic survival outcomes may be due to the lack of class 4 and 5 patients numbers (see Table 7.6). From Figure 5.6, those with CK14 expression above 150 have been observed to have poorer prognosis than those with lesser CK14 expression in the 1076 group, but this trend was not found simply due to lack of patient numbers with high CK14 expressions. From Figure 5.6, there is only one patient with CK14 $>$ 150 expression in the 413 group that satisfy the conditions to be included in the survival analysis. Therefore, the more optimistic survival outcomes in classes 4 and 5 in Figure 5.4.

Examination of the classification of the remaining 413 patients revealed similar distribution of NPI values by class as found by Soria *et al.* [141] which, not only supported earlier claims of NPI providing discriminant information, but also showed that the framework was capable of accurately classifying the 413 patients. Despite these patients previously belonged to mixed classes [141], the new classifications of the 413 patients showed characteristics consistent with those by Soria *et al.* [141]. Furthermore, the distinction between the three main classes found in the survival analysis of the classified breast cancer types for the 413 patients using EKKZ and ENB not only showed an association between the survival and breast



(a) The 1076 group (b) The 413 group

Figure 5.6: Survival curves of patients in the 1076 group and in the 413 unlabelled group based on CK14 expressions.

cancer types, but it also supported Soria's classification. In the analysis of the biomarkers, their distinct values (shown in the cluster centres) and distributions (comparisons of boxplots with Soria's classification) verified the framework as well as the importance of the protein biomarkers in characterising the classes and discriminating between them. The analysis with clinical information such as age, grade, NPI and survival showed significant associations between the biological classes and these clinical information, which can help provide support to a more accurate prognosis.

Caution needs to be exercised in trying not to "force" patients to belong to a class by choosing solutions from one technique. It is the consistency and thus, the stability of the solution from different techniques that is sought after. Thus, two approaches have been used with ssFCM (with investigation in α settings) for this framework and have shown that both approaches highly agreed with each other. More importantly, using ENB in the framework, accurate classification have been produced (based on evaluation from the biplots, boxplots and survival curves) using 15 out of the 25 features. In a separate study [100], various analysis from EKKZ

classification of the 413 patients were compared with those of ssFCM-M-KKZ (Mahalanobis ssFCM with KKZ) and more resemblance with Soria's classification was found with the former methodology.

5.6 Summary

The integrated framework incorporated initialisation techniques and/or feature selection technique into the ssFCM framework. Using EKKZ and ENB, the same six classes of breast cancer types as Soria's classification can be identified for the 413 patients. This evaluation showed that the framework was able to accurately classify the NTBC dataset. Furthermore, this confirmed the six subgroups identified by Soria *et al.* [141] and addressed the issue of stability of their classification.

By classifying the 413 patients based on labelled data of the 663 patients using ssFCM methodologies and performing different analysis on the classification result, the framework have been demonstrated to provide decision making support for clinicians. The distribution of biomarkers by class presented on the boxplots showed similar key characteristics of the six classes by Soria *et al.* The biplots showed the clusters generated from the classified patients' breast cancer type were located similarly to Soria's classification. Based on comparisons of boxplots of biomarkers, NPI values and biplots, EKKZ and ENB produced classifications of patients (previously deemed mixed classified) that closely resemble Soria's classification.

Using the ssFCM methodologies with high contribution from supervised learning using $\alpha = 30$ and with 100% labelled data was observed to produced classification which were not as clinically relevant in terms of the survival curves as with using $\alpha = N/M$ setting. The high contribution from supervised learning may not be suitable in this case as it imposes strict cluster borders. Furthermore, the 413 patients are located at the borders of clusters.

In addition, from a clinical point of view, it is hoped that a more accurate model have been provided for the prediction of breast cancer types for new patients, using both classification from Soria *et al.* and classification of the remaining 413 patients, which previously belonged to a mixture of classes, that can further help support decision making.

Emphasis has to be made on the importance in exercising caution so as not to "force" patients to belong to a class by choosing solutions from one technique [43]. Rather, the consistency and thus, stability of the solution from different techniques should be maintained, a fundamental clustering issue raised by Jain [84]. This motivated the next study on identifying the six stable subgroups by reproducing them using ssFCM and unsupervised clustering techniques.

6 Finding Stable Subgroups Using Reduced Sets Of Protein Biomarkers

The objective of this chapter is to identify stable and clinically-useful breast cancer subgroups using the integrated framework, previously introduced in Chapter 5, through investigation of the two reduced sets of biomarkers. The reduced sets of biomarkers to be investigated are 1) a set of 15 biomarkers identified in [98] and 2) a set of 10 biomarkers [140].

6.1 Background and motivation

The 15 features identified using ssFCM and NB-RFE in Chapter 4.2 have improved classification accuracy for classifying the 663 patients. Furthermore, when using these 15 features to test on the 413 patients, the same six subgroups which these patients are assigned to are tested to be biologically useful and clinically relevant. In [140], 10 important features have been found using an exhaustive search of the best combination based on the NB classification results. The same 10 important features were also used to identify the key clinical phenotypes of breast cancer [63]. In addition, these features are used in the generation of a linguistic rule set using a fuzzy rule induction algorithm for breast cancer classification [142]. These studies prompt the question as to whether these 10 important features will produce better subgroups, in terms of stability, than using the 15 that previously identified (in Chapter 4.2).

Using Pearson's correlation [129], the biomarkers have been checked to be not confounding factors as correlation with clinical data is found to be weakly correlated [149] with coefficient values between -0.3 to 0.3 (see Table 7.13 in the appendix on page 202. This ensures that the relationship

between the biomarkers and clinical data, if any, does not cause a false association between clinical data and the subgroups.

In this chapter, the effects of the two features sets on the identification of stable breast cancer subgroups are investigated. The two different feature sets are the 15 features described in Chapter 4.2 and the 10 features identified in [140, 132, 63]. To carry out this investigation, all 1076 patients in the NTBC dataset are clustered with the two considered features sets using ssFCM, consensus K-means (CKM) and model-based clustering via BIC (MBIC). The feature sets are evaluated based on the stability of the subgroups that are produced. The stability are measured based on agreement levels with Soria's classification using Cohen's Kappa index (κ).

The aim of this chapter is to identify the same subgroups using a relevant reduced feature set using all 1076 patients. This means that the same relevant feature set is able to produce stable subgroups using different clustering algorithms. The stability of the subgroups based on agreement levels helps ascertain whether the reduced feature set can reproduce the same subgroups as with all 25 features (biomarkers). This investigation using clustering algorithms, particularly unsupervised ones with no influence from Soria's classification, will not only identify stable subgroups, but ascertain the relevant feature set that characterises these subgroups. Furthermore, these stable subgroups have to be biologically meaningful and clinically relevant in order to be considered as clinically useful [9, 141], as previously discussed in Chapter 5.

Previously in Chapter 4.2, the importance of stability in feature sets was discussed as the consistency of the same features being selected indicates confidence and therefore, importance in these features. In this chapter, however, not only are the solutions from ssFCM evaluated, but also, the solutions from unsupervised clustering algorithms and their agreement across the different clustering algorithms. Furthermore, the evaluation of

these feature sets are based not only on comparison with Soria's classification of the 663 patients, but also the comparison with the entire dataset of 1076 patients. If high agreement could be found using completely unsupervised clustering algorithms such as KM and MBIC, this would indicate that stable subgroups have been found using the relevant feature set.

6.2 Selected algorithms

As ssFCM is able to retain all Soria's classification when used to classify the remaining 413 patients in Chapter 5, it is therefore employed in a clustering setting to identify the six clusters based on Soria's classification.

The unsupervised clustering algorithms considered are consensus k-means (CKM) and model-based clustering via BIC (MBIC). K-means is well-known for its simplicity and good clustering performance. However, it suffers from having suboptimal solutions due to different initialisations. Thus, a simple algorithm to reach a consensual solution from the different initialisations is devised. The fundamental idea is to place data patterns that are frequently in the same cluster together.

The second considered unsupervised clustering algorithm is the model-based clustering via Bayesian Information Criterion (MBIC). It is based on the Expectation Maximisation (EM) algorithm with modified BIC for model selection and has been described in detail in Chapter 2.

These two algorithms are chosen as they were found to reproduce the same six subgroups using the 663 patients data with high agreement to Soria's classification with κ values of 0.920 and 0.947 respectively when different clustering algorithms were explored [99]. Consensus Fuzzy c-Means was also explored but it was not able to find the same six subgroups and is, thus, dropped from further investigation.

6.2.1 Semi-supervised Fuzzy c-Means

For this study, the Pedrycz and Waletzky's ssFCM is employed but with several different modifications based on positive results from previous studies in Chapter 4 and 5. The Euclidean distance is used instead of the Fuzzy Mahalanobis used in the original algorithm for all ssFCM methodologies, as it produced favourable results as shown in Chapter 3.2.

To find a suitable ssFCM methodology which produces high agreement with the reduced feature set, the initial cluster centres generated by KKZ (detailed in Chapter 4.1) and an adjusted scaling parameter α value of 30 (detailed in Chapter 4.3) are experimented with. Previously in Chapter 4.3, the incorporation of KKZ generated initial cluster centres and/or scaling parameter α value of 30 in ssFCM have been observed to produce improved accuracy than using ssFCM alone.

6.2.2 Consensus k-means

A simple algorithm is devised, which is referred as CKM for short, to reach a consensus of K-means clustering solutions as follows:

1. Run K-means 5000 times to generate a pre-specified number of clusters, six in this case. The output is a 5000 $\times$ N matrix containing cluster labels $c_1, ...c_N$ for data patterns $x_1, ..., x_N$ for each run. The number of runs is chosen to be 5000 to ensure similar data patterns will belong to the same cluster for a majority of the runs.

2. For data pattern x_i, count the number of times it is in the same cluster as other data patterns, $count_{ij} = count(c_i == c_j)$, in all runs.

3. Repeat with different ϵ values until a biplot showing classes most similar to those by Soria *et al.* is produced:

   (a) For data pattern x_i, a list, l_i for $i = 1, ..., N$ is created containing other data patterns that share the same clusters for $count_{ij} > \epsilon$.

(b) If $x_j \in l_i \wedge length(l_i \cap l_j) > 20$, each list l_i is then updated by performing a union with all other lists that share common data patterns $l_i = l_i \cup l_j, j \neq i$. If the other list l_j fulfills this condition with l_i, l_j is deleted.

(c) The largest six lists are chosen as the clusters and other lists, usually with much smaller number of members are ignored.

(d) Present the six lists in a biplot.

The parameter ϵ is chosen by inspection of biplot of the six lists, which are essentially the six clusters. The choice of ϵ is arbitrary. A ϵ which produces clusters similar to Soria *et al.* is chosen because they have been shown to be biologically meaningful [141]. Furthermore, a small ϵ will increase the tendency for clusters to merge or overlap and a large ϵ will create more compact clusters at the cost of ignoring some data patterns. To avoid repetition of explanation on the MBIC algorithm, the reader can turn to Chapter 2.3.4 where it has been described.

6.3 Experimental methods

In Chapter 5, ssFCM-25 (with 25 features) achieved high classification accuracy on NTBC and can completely retain Soria's classification. For this reason, the solution from ssFCM-25 is used as the benchmark for comparison with the selected clustering algorithms' solutions.

Using ssFCM, CKM and MBIC, all 1076 patients are clustered with experimentation using two reduced feature sets, the 15 identified using NB-RFE and ssFCM based on 663 classified patients and the 10 identified based on NB classification performance described in [140]. For ssFCM, all 663 labelled data from Soria's classification are used for supervised learning. Initialisation techniques and $\alpha = 30$ are used to retain Soria's classification and increase level of agreement with solutions from ssFCM-25.

Agreement levels based on κ are compared between Soria's classification and clustering solutions, where only the same 663 patients Soria *et al.* had classified are considered. As ssFCM retains all of Soria's classification (explained in Chapter 5), agreement levels between ssFCM clustering solution using all 25 features (ssFCM-25) with the clustering solutions from ssFCM, CKM and MBIC based on clustering NTBC with the two reduced feature sets are compared. Confusion matrices based on the highest agreement levels obtained using one of the two reduced feature sets are shown for each clustering algorithm. To show where the disagreement occurs with respect to individual subgroups (classes), sensitivity and specificity measures [7] are used. Sensitivity measures the rate of true positives and specificity measures the rate of true negative. The confusion matrix, Cohen's κ Index and sensitivity and specificity measures are implemented using `confusionMatrix` from the `caret` R package [95].

Further analysis is conducted on the ssFCM clustering solution to determine that the subgroups generated from this investigation are stable, biologically useful and clinically relevant. The analysis are based on agreement measures, biological evaluation and clinical evaluation respectively.

The work is extended to find the same stable seven subgroups identified by Green *et al.* [63], where the HER2 group represented as class 6 is split into two subgroups, HER2/ER+ and HER2/ER-. These are manually split where class/cluster 6 patients with ER expression of more than zero are retained and those with zero ER expression are assigned to class/cluster 7. The biological usefulness and clinical relevance of these subgroups are evaluated using agreement measures and clinical evaluation respectively.

6.4 Results

Table 6.1 on page 147 shows the agreement levels of the various clustering methods using the 15 features identified using NB-RFE and ssFCM and

Table 6.1: Agreement levels using Cohen's κ Index between ssFCM, CKM and MBIC with Soria *et al.* classification [141] and with ssFCM-25.

Method	Soria's classification (663)	ssFCM-25 (1076)
ssFCM-15	1	0.977
ssFCM-KKZ-15	1	0.976
ssFCM-KKZ-15-alpha=30	1	0.968
ssFCM-10	1	0.873
ssFCM-KKZ-10	1	0.874
ssFCM-KKZ-10-alpha=30	1	0.881
CKM-25	0.693	0.597
CKM-15	0.763	0.727
CKM-10	0.755	0.650
MBIC-25	0.830	0.765
MBIC-10	0.769	0.625

10 features identified in [140]. To indicate which reduced feature set is used, the number of features with the clustering method as well as any other methodologies used are indicated. The number in brackets indicate which patient population the comparison is being made with; either the 663 patients which Soria *et al.* classified [141] or all 1076 patients.

Figure 7.2 on page 201 (in the appendix) shows that CKM and MBIC with the 15 features did not produce similar subgroups as those identified by Soria and his colleagues. The difference between the subgroups made it difficult to align their labels for direct comparison, particularly MBIC-15. Thus, only biplots comparisons are presented. Although CKM-15 has moderately high agreement with ssFCM-25 in Table 6.1, the biplot shows that CKM-15 cannot distinguish between subgroups representing class 4 and class 5.

ssFCM-15 highly agree with ssFCM-25. This is expected as the 15 features found were based on Soria's classification, whose labelled data was also used in ssFCM. All ssFCM methodologies (using 10, 15 or 25 features) highly agree with each other as they clustered based on Soria's classification. Although labels from Soria's classification were used, high agreement (> 0.87) was maintained using ssFCM methodologies with 10 features. Agreement increased with the adoption of KKZ and $\alpha = 30$.

Table 6.2: Confusion matrices between clustering solutions from ssFCM-25 and ssFCM-KKZ-10-alpha30, CKM and MBIC. The low sensitivity value is italicised.

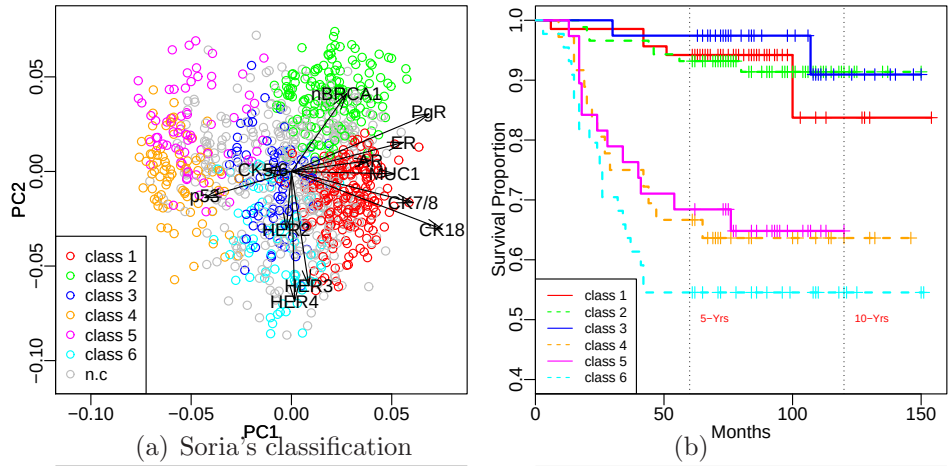
	1	2	3	4	5	6	total
ssFCM-KKZ-10-alpha30	ssFCM-25						
1	275	12	7	0	0	8	302
2	7	241	4	0	4	4	260
3	15	5	125	0	3	3	151
4	0	1	2	92	0	6	101
5	0	4	1	1	119	5	130
6	4	2	1	2	2	121	132
Total	301	265	140	95	128	147	1076
Sensitivity	0.914	0.909	0.893	0.968	0.930	0.823	
Specificity	0.965	0.977	0.972	0.991	0.988	0.988	
P-value	<0.01						
CKM-10	ssFCM-25						
1	186	2	33	0	1	1	223
2	13	208	8	0	4	0	233
3	98	49	68	0	15	25	255
4	0	1	1	87	3	5	97
5	0	1	22	4	94	2	123
6	2	1	2	2	2	102	111
o.c	2	3	6	2	9	12	34
Total	301	265	140	95	128	147	1076
Sensitivity	0.622	0.794	*0.507*	0.935	0.790	0.756	
Specificity	0.950	0.968	0.794	0.989	0.969	0.990	
P-value	<0.01						
MBIC-10	ssFCM-25						
1	172	5	24	0	1	1	203
2	5	182	7	0	5	0	199
3	119	72	98	0	17	30	336
4	0	1	3	92	2	15	113
5	0	0	4	0	100	2	106
6	5	5	4	3	3	99	119
Total	301	265	140	95	128	147	1076
Sensitivity	*0.571*	0.687	0.700	0.968	0.781	0.673	
Specificity	0.960	0.979	0.746	0.979	0.994	0.978	
P-value	<0.01						

Using CKM-10 and MBIC-10, moderate agreement with Soria's classification and ssFCM-25 was obtained with $\kappa > 0.6$. This indicates that more stable subgroups were generated using the 10 features than using the 15 features in Chapter 4.2. For CKM-10, the agreement level is higher than CKM-25. Due to the higher agreement of clustering solutions using 10 features than 15, further analysis was conducted on the clustering results obtained from using the reduced set of 10 features [140].

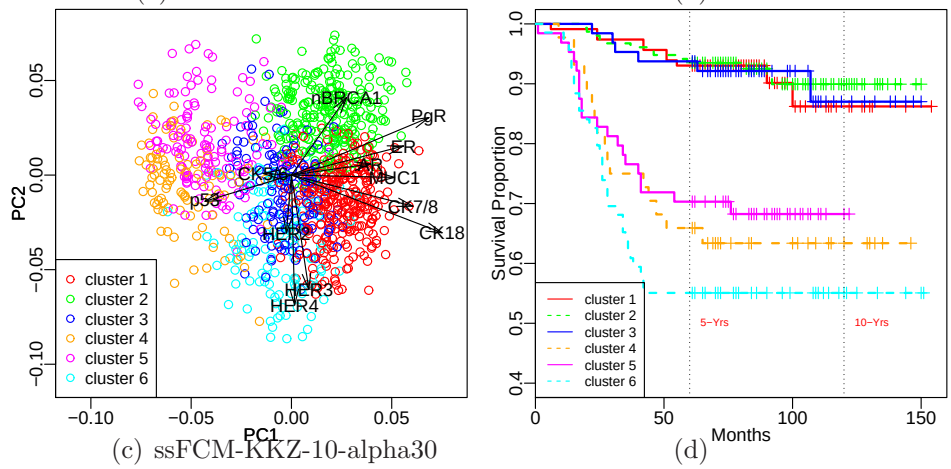
Table 6.2 on page 148 shows the confusion matrices between solutions from ssFCM-25 and ssFCM-KKZ-10-alpha30, CKM-10 and MBIC-10. Comparison between ssFCM-25 and ssFCM-KKZ-10-alpha30 shows that there is only small disagreements between the three main groups, and the disagreements within the main groups between their respective subgroups are considered small. Based on the confusion matrices and the sensitivity and specificity measures, it was observed that disagreements tended to occur within clusters 1 or 3 where there was low sensitivity (in italics) using CKM-10 and MBIC-10. Nevertheless, they both achieved average sensitivity of above 0.7 and specificity of above 0.9. The p-value indicates whether the overall accuracy rate is greater than the rate of the largest class.

Figure 6.1 on page 150 shows the biplots and survival curves using clustering solution from Soria's classification and clustering solution from ssFCM methodology, CKM and MBIC using 10 important features. Using the 10 features with the three different clustering methodologies, identical subgroups (as shown on the biplots) could be identified. Furthermore, their survival curves show clinical relevance in terms of overall survival outcome. The separability between the six survival curves which corresponds to their biological subgroups also reflects the three main breast cancer groups and six subgroups, similar to Soria's classification. These observations (from the biplots and survival curves) indicate the stability of the subgroups.

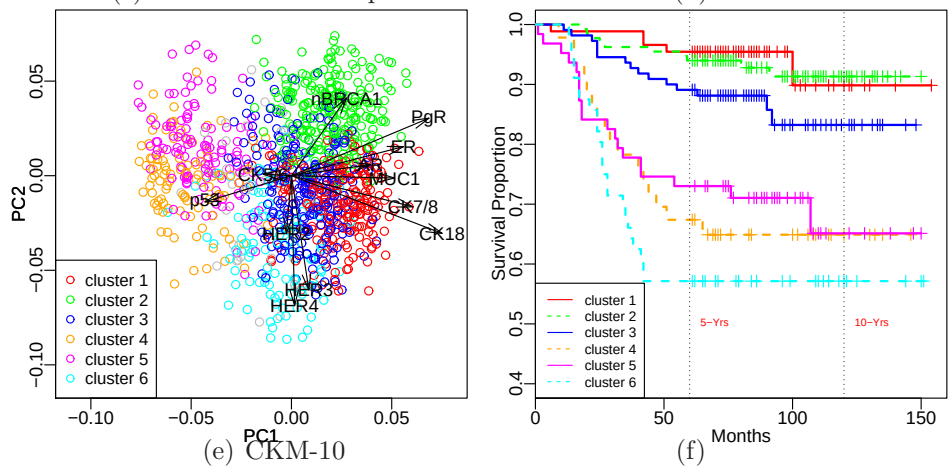
As the 10 features produced stable subgroups using different clustering algorithms and ssFCM was able to retain Soria's classification completely, clinical association of the subgroups found by ssFCM-KKZ-10-alpha30 is presented in Table 6.3 on page 151. Based on Cramer's V and, presented in brackets, p-values (Pearson's chi-squared test of independence), significant association with clinical parameters that were not involved in clustering was found. The clinical association between six subgroups identified by other clustering methods were compared in Table 6.4. The subgroups identi-



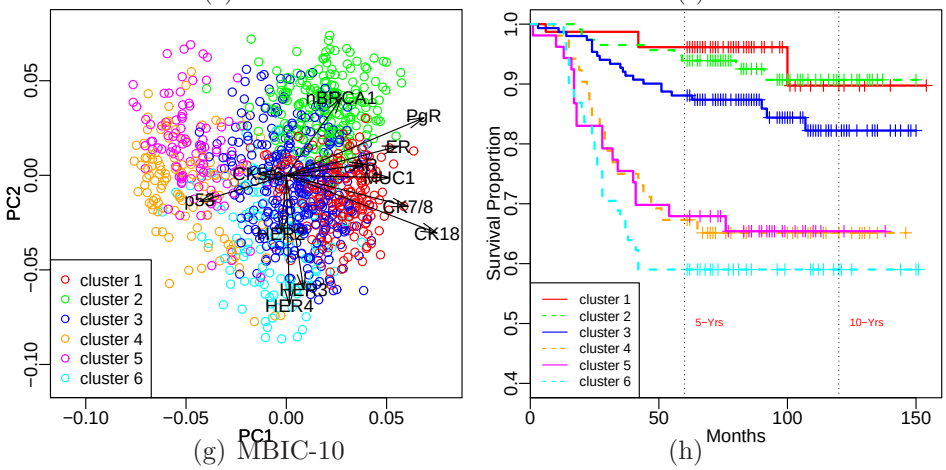
(a) Soria's classification (b)



(c) ssFCM-KKZ-10-alpha30 (d)



(e) CKM-10 (f)



(g) MBIC-10 (h)

Figure 6.1: Biplots based on Soria's classification [141](a) and clustering 1076 patients using the 10 important features by ssFCM methodology in (c), by CKM in (e) and by MBIC in (g) and their respective survival curves beside them.

Table 6.3: Association between breast cancer biological clusters from ssFCM-KKZ-10-alpha30 and clinical parameters. The values in italics denote the Cramer's V coefficient and p-values (based on Pearson's chi-squared test of independence) [2] are shown in brackets. p-values of <0.01 indicates there is association between the subgroups and clinical data.

Parameter *Cramer's V*	cla.1	cla.2	cla.3	cla.4	cla.5	cla.6
Age 0.13 (<0.01)						
⩽35	12	6	5	12	8	5
35<Age⩽45	150	123	92	30	47	53
45<Age⩽55	50	41	20	27	35	29
>55	90	90	34	32	40	45
Total	302	260	151	101	130	132
Grade 0.42 (<0.01)						
1	71	70	15	0	3	1
2	129	142	37	2	13	20
3	101	48	99	99	114	111
Total	301	260	151	101	130	132
Size 0.12 (<0.01)						
⩽1.5cm	110	109	42	19	31	29
1.5cm<Size⩽2cm	31	19	21	9	24	28
2cm<Size⩽2.5cm	78	71	43	28	35	37
2.5cm<Size⩽3cm	52	44	25	24	27	23
<3cm	31	17	20	21	13	15
Total	302	260	151	101	130	132
Stage 0.14 (<0.01)						
1	190	174	75	60	94	61
2	96	67	63	28	25	53
3	16	18	13	13	11	16
Total	302	259	151	101	130	130
Death 0.26 (<0.01)						
No	283	244	139	81	105	99
Yes	10	13	6	16	20	31
Total	293	257	145	97	125	130
NPI 0.23 (<0.01)						
⩽ 2.4 (EPG)	51	47	9	0	1	3
2.4<NPI⩽3.4 (GPG)	74	93	21	1	8	9
3.4<NPI⩽4.4 (MPG1)	78	60	38	35	44	37
4.4<NPI⩽5.4 (MPG2)	62	39	46	35	55	38
<5.4(PPG)	37	21	37	30	22	45
Total	302	260	151	101	130	132

Table 6.4: Clinical evaluation by association between clinical parameters and classification based on Soria *et al.* [141], ssFCM-KKZ-10-alpha30 (SKA-10 for short), CKM-10 and MBIC-10 measured by Cramer's V and their respective p-values (calculated using [2]) in brackets.

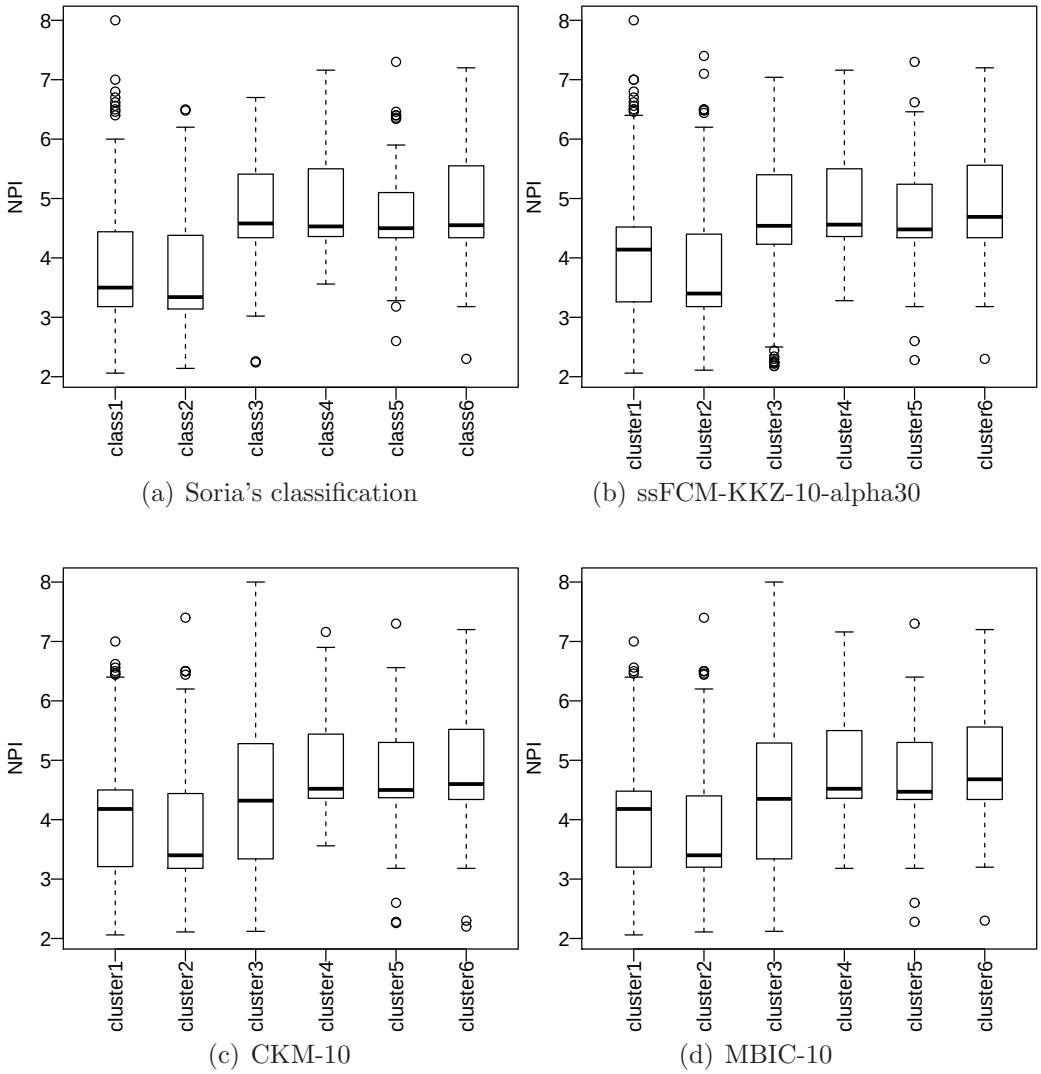
Parameter	Soria *et al.*[141]	SKA-10	CKM-10	MBIC-10
Age	0.15 (*)	0.13 (*)	0.15 (*)	0.14 (*)
Grade	0.47 (*)	0.42 (*)	0.40 (*)	0.40 (*)
Size	0.15 (*)	0.12 (*)	0.13 (*)	0.12 (*)
Stage	0.15 (*)	0.14 (*)	0.10 (0.029)	0.11 (*)
NPI	0.26 (*)	0.23 (*)	0.22 (*)	0.21 (*)
Death	0.30 (*)	0.26 (*)	0.24 (*)	0.23 (*)

* p <0.01

fied by ssFCM-KKZ-10-alpha30 have clinical associations that are competitive with Soria's classification. Higher clinical associations were found for subgroups by ssFCM-KKZ-10-alpha30 than by CKM-10 and MBIC-10 for grade, stage, NPI and death. Note that Soria's classification considers only 663 patients, while subgroups from ssFCM-KKZ-10-alpha30, CKM-10 and MBIC-10 consider all 1076 patients.

Figure 6.2 on page 153 shows the NPI distribution of each subgroup based on the different clustering algorithms. Subgroups from ssFCM-KKZ-10-alpha30 produced similar NPI distributions as Soria's classification. While subgroups from CKM-10 and MBIC-10 have similar NPI distributions as Soria's classification, their cluster 3 have a higher NPI dispersion than those from Soria's classification and ssFCM-KKZ-10-alpha30.

Table 6.5 on page 153 shows the agreement between the different clustering solutions using 10 and the original 25 features. As compared to Table 4 in [141], which shows the agreement between clustering solutions of HC (agglomerative), ART, KM and PAM, the clustering solutions obtained here have higher agreement with each other using the reduced feature set. A slightly higher agreement was found when the HER2 group was split into two. Agreement between clustering solutions of ssFCM and CKM and between CKM and MBIC increased when using 10 features. But, MBIC has higher agreement with ssFCM when using 25 features.



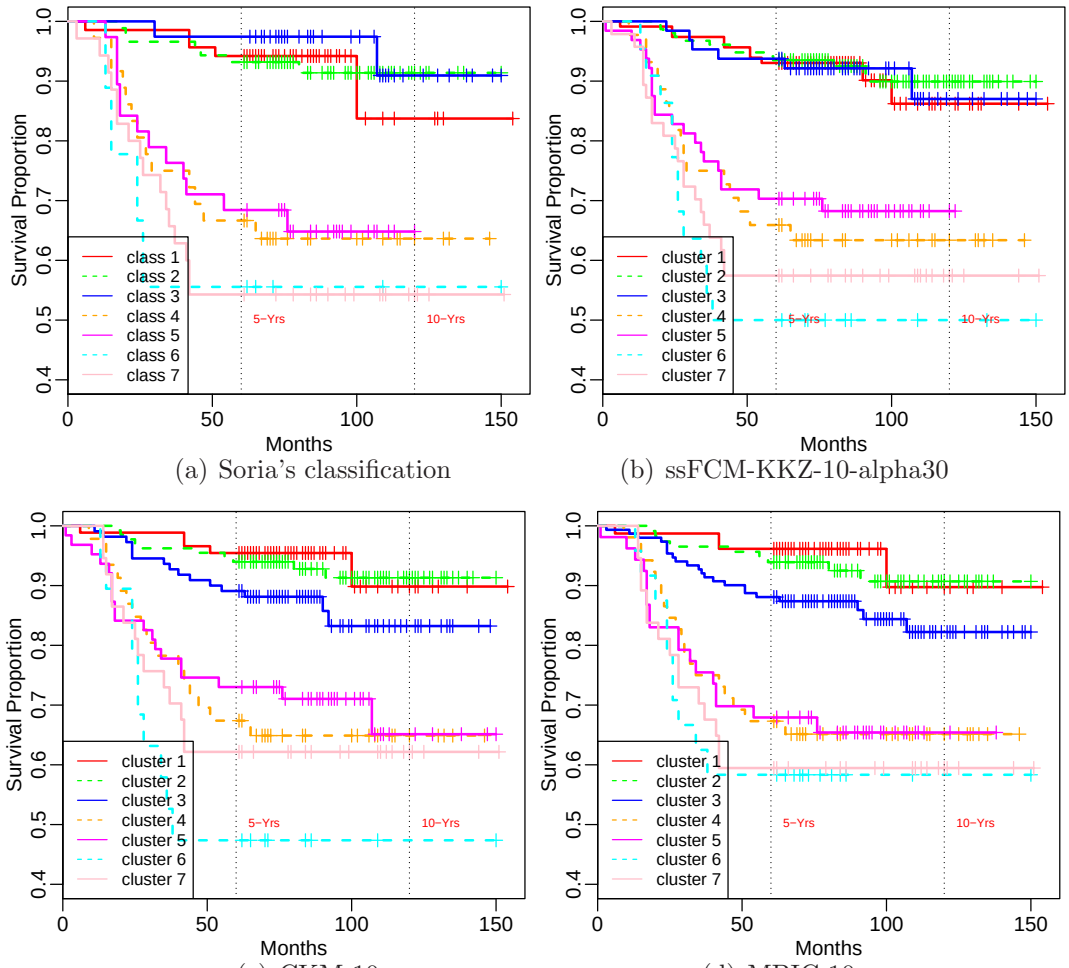
(a) Soria's classification (b) ssFCM-KKZ-10-alpha30

(c) CKM-10 (d) MBIC-10

Figure 6.2: NPI distribution based on clustering solutions with 10 features.

Table 6.5: Agreement between clustering solutions with 6 and 7 subgroups using Cohen's κ Index. SKA-10 is used as a short form for ssFCM-KKZ-10-alpha30. Agreement level is generally higher using 10 features than using all 25.

6 subgroups			7 subgroups		
	SKA-10	CKM-10		SKA-10	CKM-10
CKM-10	0.699		CKM-10	0.701	
MBIC-10	0.674	0.860	MBIC-10	0.676	0.861
	ssFCM-25	CKM-25		ssFCM-25	CKM-25
CKM-25	0.587		CKM-25	0.600	
MBIC-25	0.765	0.590	MBIC-25	0.767	0.592



(a) Soria's classification (b) ssFCM-KKZ-10-alpha30

(c) CKM-10 (d) MBIC-10

Figure 6.3: Survival curves based on 7 subgroups identified from Soria's classification [141] (a) and using the 10 important features by ssFCM methodology in (b), by CKM in (c) and by MBIC in (d).

Figure 6.3 shows the survival curves based on seven subgroups identified by the clustering algorithms with HER2 group divided into two. The survival curve based on the new subgroup is presented in pink. Survival curves based on subgroups by Soria's classification and MBIC-10 distinctively maintain both the 3 main groups and their respective subgroups. For CKM-10, distinction between survival curves for Basal and HER2 group is not clear while for ssFCM-KKZ-10-alpha30, the distinction between these two main groups are not as clear as Soria's or MBIC-10's. For survival curve analysis based on the three main groups, please refer to Figure 7.3 in the appendix on page 203.

Table 6.6 on page 155 shows the survival curve differences based on 7 subgroups and 3 main groups from ssFCM-KKZ-10-alpha30 using G-rho

Table 6.6: Differences in survival curves in ssFCM-KKZ-10-alpha30 using Kaplan-Meier p-values.

ssFCM-KKZ-10-alpha30 7 subgroups *							3 main groups *		
Cluster	1	2	3	4	5	6	Cluster	1	2
2	0.743						2	*	
3	0.989	0.785					3	*	0.104
4	*	*	*						
5	*	*	0.001	0.632					
6	*	*	*	0.260	0.105				
7	*	*	*	0.466	0.214	0.575			

* p <0.001

family of tests proposed by Harrington and Fleming [73]. The test determines whether there is a difference between one or more survival curves where a p-value of less than 0.05 means that they are different. It was implemented using the `survdiff` function from the `survival` R package [150]. Survival curves differ significantly between Luminals (clusters 1-3) and the other 2 groups, Basals (clusters 4 and 5) and HER2 (clusters 4 and 5), showing poorer prognosis in the more aggressive groups Basals and HER2 [3]. This is also reflected in the survival curve differences tables for Soria's classification, CKM-10 and MBIC-10 in Tables 7.14, 7.15 and 7.16 respectively, found in the appendix on pages 202 to 204. More importantly, the separability between the survival curves which reflects the biological subgrouping performed by the clustering algorithms indicates clinical relevance, which means that these subgroups found were also clinically useful.

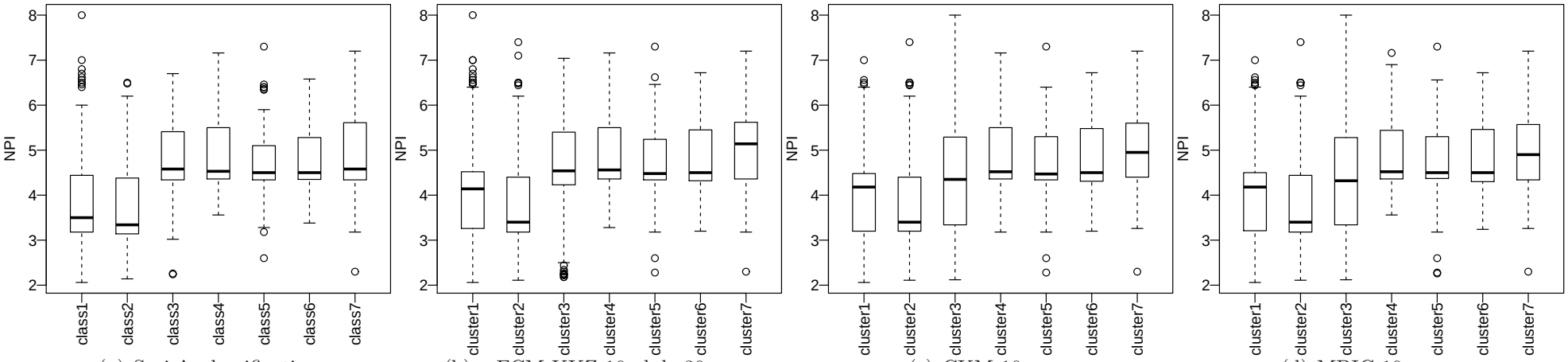
Table 6.7 on page 156 shows the association between biological subgroups and clinical parameters based on the different clustering algorithms used and the number of subgroups found. The clinical association levels for 6 and 7 subgroups were similar and thus, similar observations as previously discussed were found.

Figure 6.4 on page 156 shows the NPI distributions of the seven subgroups based on the different clustering algorithms. Apart from subgroup 3, similar NPI distributions were found on all other subgroups using the different clustering algorithms. Furthermore, similar NPI distributions for all seven subgroups as those in [63] were found using CKM-10 and MBIC-10.

Table 6.7: Clinical evaluation by association between clinical parameters and classification based on Soria *et al.* [141], ssFCM-KKZ-10-alpha30, CKM-10 and MBIC-10 measured by Cramer's V and their respective p-values in brackets based on 6 and 7 subgroups (SG).

	Soria *et al.*[141]		ssFCM-KKZ-10-alpha30		CKM-10		MBIC-10	
Parameters	6 SG	7 SG	6 SG	7 SG	6 SG	7 SG	6 SG	7 SG
Age	0.15 (*)	0.16 (*)	0.13 (*)	0.13 (*)	0.15 (*)	0.16 (*)	0.14 (*)	0.14 (*)
Grade	0.47 (*)	0.47 (*)	0.42 (*)	0.42 (*)	0.40 (*)	0.40 (*)	0.40 (*)	0.40 (*)
Size	0.15 (*)	0.15 (*)	0.12 (*)	0.12 (*)	0.13 (*)	0.13 (*)	0.12 (*)	0.12 (*)
Stage	0.15 (*)	0.16 (*)	0.14 (*)	0.14 (*)	0.10 (0.029)	0.10 (0.036)	0.11 (*)	0.11 (0.011)
NPI	0.26 (*)	0.26 (*)	0.23 (*)	0.24 (*)	0.22 (*)	0.22 (*)	0.21 (*)	0.22 (*)
Death	0.30 (*)	0.30 (*)	0.26 (*)	0.26 (*)	0.24 (*)	0.25 (*)	0.23 (*)	0.23 (*)

* p <0.01



(a) Soria's classification (b) ssFCM-KKZ-10-alpha30 (c) CKM-10 (d) MBIC-10

Figure 6.4: Boxplots showing NPI distribution for 7 subgroups.

6.5 Discussion

Using the 10 features identified in [140], stable subgroups were found using the three clustering algorithms, ssFCM-KKZ-10-alpha30, CKM-10 and MBIC-10, with agreement of above 0.6 when compared with ssFCM-25. Based on the clinical evaluation using association measure, survival analysis and NPI boxplot analysis, the subgroups identified using all three clustering algorithms were found to be clinically relevant. The ssFCM subgroups had the highest association with grade, stage, NPI and death in comparison with subgroups identified by CKM and MBIC.

Further comparison were made with the 7 subgroups identified by Green *et al.* [63], the latest development of subgroup identification in the NTBC dataset. The HER2 group was manually split into two such that those with ER > 0 belong to the HER2/ER+ group (class/cluster 6) and those with ER $= 0$ belong to the HER2/ER- group (class/cluster 7). On comparison of clinical association with subgroups by Green *et al.* [63], competitive association levels were found with the subgroups found using the ssFCM framework. NPI distributions of subgroups based on CKM-10 and MBIC-10 were similar to those of Green *et al.* [63]. This further ascertain the importance of the 10-feature set in identifying stable subgroups using different clustering techniques and methodologies.

Based on the increased agreement observed between clustering solutions from ssFCM, CKM and MBIC as compared to agreement levels of clustering solutions in [141], this suggests that the subgroups from different clustering algorithms stabilise with a suitable, reduced feature set. The significant increase in agreement between CKM and MBIC with 10 features warrants further investigation between agreeing solutions, which may produce clearer distinction between Basal and HER2 group for the 7 subgroups of CKM-10.

The 15 features found using ssFCM and NB-RFE are useful for achieving high classification accuracy when assigning new patients to classes, but

from this study, they were not useful for finding stable subgroups when used with unsupervised clustering algorithms. This may be due to the following two reasons. The 15 features were identified based on Soria's classification of the 663 patients, not all 1076 patients. There may be redundant features still remaining in the 15 features and further feature selection from the 15 feature set is required.

The increased stability of subgroups generated by clustering algorithms using a reduced panel of protein biomarkers opened up two research questions, which to the best of our knowledge, are currently not answered:

1. Can feature selection help clustering algorithms produce more stable clusters?

2. Can the stability of clusters be an evaluation criteria for unsupervised feature selection using clustering algorithms to find relevant features?

6.6 Summary

In this chapter, clustering was performed using ssFCM, CKM and MBIC with experimentation on two different feature sets, one containing 10 features and the other 15 features. Using 15 features, ssFCm achieved high agreement with ssFCM-25. However, very poor agreement were found using unsupervised clustering, indicating that using the 15 features, the subgroups found were unstable. Using the 10 features on three different clustering algorithms, stable breast cancer subgroups that are biologically useful and clinically relevant were found.

The six subgroups found using 10 features were split into seven to make comparison with other subgroups found in [63]. Competitive clinical association and similar NPI distributions in the seven subgroups were found. This further confirms the importance of the 10 features in identifying stable subgroups using different clustering algorithms and other methodologies for the breast cancer data.

7 Conclusion

This thesis focused on the development of ssFCM-based techniques for application on a real biomedical data. The incorporation of the three approaches initialisation techniques, feature selection and adjustment of scaling parameter α into a ssFCM framework, resulting in a novel ssFCM-based framework, has been demonstrated to address the problem of initialisation sensitivity and improve clustering or classification results, thereby fulfilling the first main objective. The framework has been applied to classify new patients in the NTBC dataset into the six subgroups previously identified by Soria *et al.* [141]. Furthermore, the framework is applied on the NTBC dataset with a reduced panel of 10 biomarkers to identify stable breast cancer subgroups (both six and seven subgroups). Agreement with Cohen's κ index of near 0.7 between solutions from the framework and two clustering algorithms indicate high stability in the identified subgroups. Analysis showed that these subgroups are identical to Soria *et al.* [141]. The framework has demonstrated success in being applied as a classification and clustering tool to assist clinicians in decision making, fulfilling the second main objective.

In this chapter, the thesis is summarised by describing the contributions to knowledge derived from this research and the limitations based on the extent to which the sub-objectives have not fulfilled. Approaches to tackle the limitations are discussed in the future work section. Furthermore, the publications based on this research are also listed.

7.1 Contributions to knowledge

In this section, the contributions to knowledge derived from this research work and limitations, if any, are discussed. Publications that have been generated from this research are listed and the potential avenues for future exploration are described.

A comparison between ssFCM algorithms on popular datasets

A comparison between different ssFCM algorithms was conducted to select the most suitable to be applied on a real biomedical dataset. Different ssFCM algorithms have previously been evaluated on different datasets using various amounts of labelled patterns, which makes fair comparison difficult. This prompted the study into a comparison of four simple but good-performing distance-based semi-supervised Fuzzy c-Means (FCM) algorithms proposed by Pedrycz and Waletzky [121], Zhang *et al.* [165], Li *et al.* [101] and Endo *et al.* [47]. The four algorithms were analysed and applied over five popular UCI datasets. Scale differences of dimensions in the dataset, distance metrics, objective functions and quality of labelled patterns have been observed to affect classification results. For most of the dataset tested on, ssFCMs with Fuzzy Mahalanobis distance were found to be most suitable because they are scale-invariant given that there are scale differences in the dimensions of the datasets, as compared to Euclidean and kernel-based distances. By arbitrarily reducing the dimensions in the dataset, WOBC in this case, improved ssFCM results were obtained. This is one motivation towards the investigation of feature selection techniques to improve ssFCM classification (as was conducted and described in Chapter 4.2). In Li-08 [101], only labelled data are used to update the cluster centre which makes the algorithm highly dependent on the initial membership values of labelled patterns, which if contain errors, can negatively

impact the classification results. This brings the importance of scaling (balance) parameters like α and a. In Zhang-04 [165], only memberships of unlabelled patterns are updated and thus, memberships of labelled patterns are never updated, nor improved. Endo-09 [47] performed the least favourably due to the use of Euclidean distance on scale-variant datasets which faced the 'curse of dimensionality' problem. Based on these observations, further investigation in Pedrycz-97 [121] is conducted for application on a real biomedical dataset, the NTBC dataset.

It was observed that accuracy did not always increase with amount of labelled data, indicating that some labelled data may not be good example. The techniques for selecting "good" labelled data have not been explored because there may be a danger of selecting labelled data that creates clusters that are thought to be correct, rather than let the data and algorithm determine this. Moreover, given that labelled data are limited, all labelled data are considered valuable in providing more information regarding the structure within the data. By filtering out "bad" labelled data, finding anomalies which insights can be gained from may be hindered. Furthermore, these labelled data "good" or "bad" are representatives of the population. Besides, there is a general trend of increasing accuracy despite slight fluctuations in accuracy as amount of labelled data increases. Based on these reasons, all available labelled data are used rather than devising a selection mechanism for labelled data in the NTBC dataset. This fulfills sub-objective 1a of this research.

So far, the comparison was performed on four distance-based ssFCMs. Pairwise-constrained ssFCMs [64, 59] are also distance-based ssFCMs but, is not covered in this research. This is because additional parameters are at play in these type of ssFCMs, which require further study.

Demonstration of ssFCM robustness using evaluation techniques

The ssFCM was evaluated using accuracy rate, κ and NMI in two settings; clustering and cross-validation in Chapter 3.2. The results generated using the three evaluation techniques were consistent with each other, demonstrating robustness of the algorithm. Thus, κ and NMI are dropped in proceeding investigations. Furthermore, the relative accuracy (suggesting improvement or otherwise) between different techniques or methodologies were more of interest than the actual accuracy when testing on other datasets, which evaluation using accuracy rate suffices. This fulfills sub-objective 1b of this research.

Investigation of distance metrics in ssFCM for breast cancer data

The investigation in distance metric was prompted by two factors. First, a preliminary study on the application of ssFCM for breast cancer classification [99] showed low agreement with Soria's classification [141] using ssFCM algorithms with Fuzzy Mahalanobis distance on NTBC. Secondly, based on literatures [85, 84] and on experimental results from the preliminary studies (in Chapter 3.1) on comparing different ssFCM on different UCI datasets, it is clear that the choice of distance metric strongly affects the performance of ssFCM for different datasets. This led to the exploration of different distance metrics to find one that best represent the NTBC dataset. Unexpectedly, the Fuzzy Mahalanobis distance commonly used in many ssFCM algorithms did not perform well for NTBC.

Soria's classification [141] was used as a benchmark because the subgroups identified in [141] were shown to be biologically meaningful and had significant clinical association. In Chapter 3.2, breast cancer classifications based on using Euclidean, Mahalanobis, Fuzzy Mahalanobis and kernel-based distances in ssFCM (Pedrycz-97) were compared. The highest agreement was obtained from Euclidean ssFCM. Interestingly, the Maha-

lanobis ssFCM was found to perform better than the Fuzzy Mahalanobis one on the NTBC dataset. Similar experiments were ran on other UCI datasets (not included in this thesis) where different distance metrics were found to be suitable for different dataset. This suggests the importance of experimenting an algorithm with different distance metrics to determine one most suitable for a dataset. In many ssFCM literatures, the most commonly used distance metric is the Fuzzy Mahalanobis. However, it was observed that Fuzzy Mahalanobis does not always produce better results than the original Mahalanobis on all datasets but depends on the dataset. This fulfills sub-objective 2a of this research.

So far, four distance measures were explored as they were found to be most commonly used in classification or clustering of numerical data. One limitation is that the selection of suitable distance metric for a dataset is through trial and error as observed in Chapter 3.2. A distance metric that is adaptive to the relationship in the data may improve clustering or classification results.

Improving ssFCM classification of NTBC and UCI datasets using initialisation techniques.

An investigation in initialisation techniques SCS, KKZ and CE, with ssFCM on the NTBC and three UCI datasets was conducted to improve classification, which to the best of our knowledge has never been done before. ssFCM with KKZ initialisation was demonstrated to improve classification accuracy for the NTBC dataset. It was observed that initialisation techniques can improve ssFCM classification particularly when availability of labelled data is very low. The initial cluster centres provided additional information for ssFCM when labelled data is lacking. This indicates that KKZ initialised cluster centres based on information from the dataset can provide supplementary supervision in addition to labelled data to assist

ssFCM classification. The biplot analysis of the KKZ cluster centres revealed that they are located at the edge of clusters. Interestingly, despite the clusters centres found by KKZ being at the edge of clusters, improvement in accuracy is still achieved, showing the importance of employing some heuristics (KKZ, in this case) to give ssFCM some clues of where the clusters initially are, even though they may not be at the exact centre of the clusters. Indeed, ssFCM classification of the NTBC dataset has already achieved high accuracy of above 90%. However, the incorporation of initialisation technique added information that could help achieve a higher accuracy, which is beneficial to classification of future breast cancer patients and to the identification of useful clusters. Although the improvement in classification accuracy was small for NTBC, this is considered important as accuracy is critical when dealing with biomedical data. Furthermore, ssFCM methodology with initialisation also showed improvement in popular UCI datasets Iris, Wine and Pima Indian Diabetes. This contribution fulfills the sub-objectives 1c, 1e and 2a.

Improving ssFCM classification of NTBC using feature selection techniques and identification of important biomarkers.

The motivation is to produce clinically useful classification of NTBC using as few features as possible, thereby reducing cost and time in running clinical tests. Feature selection is applied with ssFCM and investigations in accuracy and stability of selected features are carried out. If a reduced feature set achieves high stability, it means that the same set is consistently being selected and this builds confidence in the feature set. For NTBC, ssFCM with the 15 features selected by NB-RFE achieved the highest average accuracy–stability combination. ssFCM with 17 features selected by NB-RFE achieved the highest average accuracy but lower stability than with 15 features.

The best 15 features are identified by choosing the highest scoring features based on their rank and frequency of belonging to a feature set that achieve 100% ssFCM classification accuracy. To the best of our knowledge, the application of NB-RFE with ssFCM as evaluation of the selected features has never before been done. While improvement in accuracy was found using ssFCM on the 15 selected features, it is crucial to test the same 15 features on different classifiers to ensure that these features are not biased to some classifiers only. Should a different classifier be used on NTBC in the future, these features may not produce accurate results. Eleven classifiers were tested using the reduced feature set of 15 features and nine of these classifiers showed improvement in accuracy compared with them using all 25 features. Apart from FHIT, these features are consistent with the overexpressed and underexpressed features presented by Soria *et al.* [141].

To demonstrate the generalisation of the ssFCM methodology with feature selection on other datasets, the methodology is applied on three UCI datasets Arrhythmia, Cardiotocography and Yeast with investigation in distance metric, accuracy and stability. Accuracy increased on all three datasets using the methodology. Unexpectedly, it was observed that ssFCM with a poor performing distance metric (and original number of features) when used with feature selection could obtained much improved accuracy, higher than that with the most suitable distance metric and original number of features. This suggests that further investigation in different distance metrics in ssFCM to determine one best for the dataset with reduced feature set is required. This finding, to the best of our knowledge, has never been reported. This contribution fulfills the sub-objectives 1d, 1e and 2a.

With the newly selected features, the hidden structure within the reduced dataset has changed and may require the search of a suitable distance metric. This increases computation requirement and the uncertainty of finding the right distance metric to produce useful clusters.

Investigation of scaling parameter α for improving ssFCM

The scaling parameter α helps maintain a balance between supervised and unsupervised learning in ssFCM. In [59] and [155], the selection of suitable α values were found to depend on the dataset and the amount of labelled data used. This was demonstrated on UCI datasets Ionosphere, Page Blocks , PID, Wine and WOBC as well as NTBC in Chapter 4.3. Based on the experimental results, three trends in accuracy are identified with respect to α and amount of labelled data for the different UCI datasets. Accuracy can be improved by either 1) increasing α irrespective of amount of labelled data, 2) increasing α with increasing amount of labelled data and 3) decreasing α with increasing amount of labelled data, depending on the dataset. The trends observed in the results suggest that accuracy can be improved when a suitable configuration for α is employed, based on prior experimentation of different α settings on the dataset. As accuracy is of great importance, further increase in accuracy obtained from suitable α configuration is desirable.

For the NTBC dataset, it was observed that increasing α increased accuracy irrespective of amount of labelled data and that $\alpha = 30$ improved classification by ssFCM and ssFCM methodologies using KKZ, 15 best features identified using ssFCM and NB-RFE and a combined methodology of KKZ and feature selection. Based on comparison between experimental results from the different ssFCM methodologies and on the amount of labelled data, a strategy is devised to use a suitable ssFCM methodology to achieve a higher accuracy.

Chapter 4.3 showed that a strategy can be devised, through prior experimentation with different α values and analysis of its effects on the result, to exploit α in ways specific to the dataset to increase accuracy of ssFCM classification. This contribution fulfills the sub-objectives 1d, 1e and 2a.

Comparing breast cancer classification between ssFCM and other classifiers

The comparison of ssFCM with other classifiers on NTBC determines how well ssFCM work in comparison with other techniques and justify the choice in ssFCM as a suitable algorithm for NTBC. So far, ssFCM has been shown to outperform 11 out of the selected 12 classifiers in classifying the NTBC dataset. With only 10% labelled data, ssFCM was found to be one of the most favourable algorithm in close competition with GLMNET. With 60% labelled data, ssFCM produced the highest accuracy of 97.84%. To the best of our knowledge, such a comparison between ssFCM and other classifiers on the NTBC was never done before. This contribution fulfills the sub-objective 2b.

So far, apart from specifying the number of classes or clusters, the classifiers are applied using default configurations. Other than for comparison, further in-depth study of their application in breast cancer classification should be considered which may improve classification and clustering results of the ssFCM framework on the entire dataset as was conducted in Chapters 5 and 6.

An integrated framework for automatic classification (post-initialisation) of new patients

An integrated framework was developed for automatically classifying the NTBC dataset into the same six subgroups using one single clustering algorithm as a tool to assist clinicians in decision making. The remaining 413 patients (belonging to mixed groups [141]) are classified using the framework and the biological and clinical characteristics of the subgroups formed by these patients, are analysed and compared with those of the 663 classified by Soria *et al.* [141]. The classification of the 413 patients exhibit similar characteristics as those in the already found six classes. ssFCM clas-

sified the 663 patients with Soria's classification fully retained. Based on clinical evaluation, clinical association was found in the six classes the 413 patients are assigned to, consistent with those reported by Soria *et al.* [141]. The resulting classification provided relevance (association) to clinical data, allowing clinicians to draw conclusions from information about the biological classes and their relevance to clinical data, hence, providing decision making support.

In doing so, a tighter integration between the clustering algorithm and application needs [84] is achieved. In classifying the remaining 413 patients, it is hoped that a more accurate model is provided for the prediction of breast cancer types for new patients that can help support decision making. This contribution fulfills the sub-objectives 2a, 2c and 2d.

When using $\alpha = 30$, ssFCM methodology with KKZ and with the 15 identified features produced highly similar classification. Although EKKZ30 and ENB30 appear not to be as clinically relevant based on their survival curves as compared to those of EKKZ and ENB, these classification are considered useful for comparing the effects $\alpha = 30$ has on the classification of the 413 patients and their eventual effect on the survival curves. Perhaps, $\alpha = 30$ is not the best configuration for classifying the 413 patients or that there is insufficient test data to demonstrate the distinct separability of the survival curves based on the subgroups assigned. Furthermore, the setting $\alpha = 30$ has been chosen based on experimentation with the 663 patients in Chapter 4.3. Further investigation is required in determining the suitable setting of α for the classification of the 413 patients. In addition, further investigation using more breast cancer data is needed.

The framework involves a two-step approach to link the subgroups to survival data, both of which are separate. First, the patients are assigned to the subgroup through classification and then survival curves based on the subgroups they belong to are drawn to establish the link. It would be

ideal if the survival data (and/or clinical data) can be incorporated into the classification process to create subgroups that are both biologically meaningful and clinically relevant directly.

While the framework is able to generate information required by clinicians for decision-making, it is currently not fully automatic. The classification results have to be fed into different scripts to generate the statistical information to serve the decision-making purpose.

In the clinical evaluation, missing data are ignored. Thus, patients with missing clinical data are not included into the analysis for establishing association between the subgroups and clinical parameters.

Identification of stable breast cancer subgroups using reduced panel of biomarkers

The ssFCM framework is also applied to identify stable breast cancer subgroups using a reduced panel of 10 biomarkers. The framework incorporates KKZ and the $\alpha = 30$ setting. The stability of the identified subgroups by ssFCM are evaluated based on agreement levels with two unsupervised clustering algorithms, CKM and MBIC. Using the 10 biomarkers on three different clustering algorithms, a moderate agreement level of above 0.6 (near 0.7) were obtained between the three clustering solutions. These agreement levels indicate that these subgroups are more stable than those in [141] using HCA, KM and ART where κ of below 0.5 were obtained. Furthermore, all 1076 patients are assigned to one of the six subgroups. The biomarker profiles for each subgroup represented in the boxplots showed biological meaningfulness and the significant clinical association showed clinical relevance of the subgroups. This study not only ascertained the importance of the 10 biomarkers but also, six stable subgroups have been identified and shown to be biologically useful, retaining all of Soria's classification, and clinically relevant, demonstrating significant clinical associations.

Further analysis on seven subgroups are conducted by manually splitting the HER2 group into two subgroups based on high and low ER expressions (HER2/ER+ and HER2/ER- subgroups respectively) and retaining the other five subgroups to form a total of seven subgroups to make comparison with the seven identified by Green *et al.* [63]. Slightly higher agreement between the ssFCM methodology with CKM and MBIC clustering solutions were found, suggesting higher stability in the seven subgroups than six. Furthermore, competitive clinical association and similar NPI distributions in the seven subgroups from ssFCM were found when compared with those by Green and his colleagues. This demonstrates that the ssFCM methodology can identify stable and clinically useful breast cancer subgroups. This study further confirms the importance of the 10 biomarkers in identifying stable subgroups (both six and seven) using various clustering algorithms. The study of identifying stable subgroups using these 10 biomarkers and ssFCM with comparison to unsupervised clustering solutions have so far not been done before. This contribution fulfills the sub-objectives 2a, 2c and 2d.

The increased stability of subgroups generated by different clustering algorithms from a reduced set of protein markers reported in Chapter 6 opens up two research questions which can bring about future technical contributions and are also highly relevant to this research objectives:

1. Can feature selection help clustering algorithms produce more stable clusters?

2. Can the stability of clusters be an evaluation criteria for unsupervised feature selection using clustering algorithms to find relevant features?

So far, only two unsupervised clustering algorithms are compared with the ssFCM framework. More unsupervised clustering algorithms should be explore to strengthen the findings and verify whether feature selection

can help clustering algorithms produce more stable clusters, as outline in research question 1. Furthermore, the exploration of other unsupervised clustering algorithms for the purpose of feature selection may help to answer research question 2.

The 15 features identified were found to produce less stable subgroups than compared with the 10 features from [140] when comparing clustering solutions from ssFCM, CKM and MBIC based on clustering the entire dataset of 1076 patients. The poor stability may be due to the fact that the best 15 features found were based on classification of the 663 patients and not on the entire dataset. Further investigation in feature selection based on data and class labels of all 1076 patients is required.

Discussion

In this research work, the ssFCM is recognised as both a clustering and classification technique. It does classification (prediction of class labels) through clustering, that is based on similarity of data. This means that the model (based on both train and test data) is updated even during testing phase in classification. However, it is not a classification technique in a machine learning sense because it does not learn the mapping between labelled data and their class labels. In classification (machine learning), the learned mapping (or model) does not change during the testing phase. At what point do we move from clustering to classification (machine learning) is an important and difficult question. This depends on how "useful" or "meaningful" are the clusters found and is not within the scope of this research. This research has used known test labels to evaluate the meaningfulness of the clusters. Where test labels are not available, association measures between clusters and other additional features in the dataset are employed as in the case of the NTBC dataset.

Based on experimental results, a simple-to-implement ssFCM by Pedrycz

and Waletzky [121] have been demonstrated to further improve clustering and classification through modifications using suitable distance metric and suitable α setting and through integration of initialisation and feature selection techniques. Based on experiments on NTBC, significant improvement was found using distance metric and feature selection, where improvement can be further increased using a suitable α setting. In the individual experiments investigating distance metric and feature selection, it was observed that the average accuracy could vary greatly between the different distance metric and feature selection technique chosen respectively. The initialisation technique with a suitable α setting was found to be particularly effective in improving ssFCM performance when labelled data are low (at below 20% labelled data). While these techniques and modifications have been widely reported and are available for use, no study has integrated these into an ssFCM framework to improve its performance. Our research have shown that a simple ssFCM can produce significant favourable results on real-world biomedical data such as the NTBC through investigation in these simple existing approaches.

In this research work, a more complex ssFCM, the pairwise-constrained ssFCM with the competitive agglomeration feature removed, have been explored. However, the user-specified parameters were found to be tricky to set and they are dependent on the dataset. As a result, we were not able to replicate the results obtained in [64] and the investigation is still ongoing. Based on this experience, the intuition is that a simple ssFCM such as Pedrycz and Waletzky [121] should first be explored and exploited using existing approaches, particularly the suitable distance metric and feature selection technique, to solve a complex problem such as clustering the breast cancer dataset before venturing into more complex approaches. Simple modifications such as using a suitable α setting can bring favourable results. This research work has taken the direction as stated by Jain [84], that is to

achieve a tighter integration between clustering algorithms and application needs. Here, a range of different modifications and approaches to ssFCM have been explored to provide improved clustering or classification result tailored to different datasets.

7.2 Future work

The proposals for future work, some of which are currently ongoing, are listed in this section. The type of contribution from each proposal is indicated in brackets whether the aspect is technical, clinical or both.

Investigation into pairwise-constrained ssFCM for breast cancer classification (technical)

So far, a modified ssFCM based on the ssFCM by Pedrycz and Waletzky [121] has been employed on NTBC where the distance metric used is Euclidean and the scaling parameter is $\alpha = 30$ with the incorporation of KKZ and reduced panel of features. Similar investigations into a different type of ssFCM, a pairwise-constrained ssFCM would help provide a deeper understanding of how distance metric, scaling parameter, initialisation techniques and feature selection affect the algorithm in comparison to the modified ssFCM. These ssFCM are of interest because they incorporate an agglomerative mechanism which do not require prior knowledge of the number of clusters. It would strengthen this research work to employ pairwise-constrained ssFCM for the purpose of determining the number of clusters for the NTBC dataset and of further analysis in the generated clusters, whether they are clinically useful or stable. This study also goes in line with the aim of using ssFCM methodologies for application in biomedical data, expanding to more complex algorithms. Furthermore, a comparative study between our integrated framework with a more complex pairwise-

constrained ssFCM help us strengthen our research work as to whether the integration of approaches and modification of a simple ssFCM can achieve as good or better results as more complex ssFCMs.

Further investigation into a more adaptive distance metric approach with ssFCM on the NTBC dataset (technical)

Distance metric learning One approach to improve the procedure of selecting suitable distance metric for a dataset is to employ distance metric learning (DML). Distance metric learning ensures that the distance relation among the training data is preserved using labelled data to indicate whether they are similar and dissimilar. By using distance metric learning, data patterns are transformed such that similar data patterns are placed closer and dissimilar ones are pushed further apart. Furthermore, distance metric learning techniques also perform feature reduction which could remedy the problem of trying to obtain improved classification results on datasets with newly selected features through the search for a suitable distance metric. But, the problem with distance metric learning, unlike feature selection, is that the output is a transformation matrix which shows no indication of which features exactly are important.

The distance metric learning (DML) techniques by Xing *et al.* [163], Weinberger *et al.* [157], Goldberger *et al.* [61] and Globerson and Roweis [60] have been applied with ssFCM on NTBC and UCI datasets Arrhythmia, Cardiotocography, Yeast, PID and WOBC (results not presented) in a CV setting. While accuracy was found to improved using DML on the UCI datasets apart from WOBC, none of the techniques improved classification accuracy for NTBC. For this reason, DML was not employed in the integrated framework and further investigation has been dropped. The reason DML has not improve ssFCM classification for the NTBC is still unclear and this investigation is currently ongoing. Further investigation may also

help to determine the types of datasets DML is suitable for or whether parameters have to be adjusted to suit the datasets.

General distance metric Another investigation to improve the selection of distance metric is the application of a general distance metric for ssFCM. The objective in applying a generalising distance metric function with ssFCM classification is to automate the search of an adaptive distance metric suitable to a dataset during the training process. A general distance metric which can intelligently consider all relevant aspects of a dataset to best represent a useful model would be ideal. One such study introduces a general distance metric that is flexible to FCM has been published [160]. Another study which is of interest is the application of Bregman divergence into ssFCM. The Bregman divergence holds a family of distance metric functions, thus, providing generalisation of these functions, which is relevant to this study of general distance metrics. Banerjee *et al.* [10] have demonstrated clustering with Bregman divergence on EM, pioneering the development of Bregman soft-clustering parametric algorithms. To apply these algorithms for identifying breast cancer subgroups, the development of a non-parametric version suitable for NTBC is needed.

Further investigation into scaling parameter for classifying the 413 patients (technical)

The adjustment of scaling parameter to suit NTBC is through trial and error and based on experimentation on the 663 patients. Perhaps, a scaling parameter that is proportional to M/N where M is the amount of labelled data and N is the total number of data patterns, is more suitable for classifying the 413 patients. Further investigation into selecting a suitable scaling parameter, through parameter tuning methods or heuristics, for classifying the 413 patients is required.

Investigation into a clustering process to produce clinically meaningful subgroups using both the survival (or clinical data) and biological data (technical and clinical)

It has been observed that the breast cancer subgroups have relevance to clinical parameters. For instance, the separability between the survival curves reflects the different levels of survival outcomes of the different subgroups and main groups. Instead of trying to establish a link between biological subgroups with survival or clinical parameters, outside of the clustering process, can stable and clinically useful subgroups be found if the biological data is linked with survival data for clustering, whether the survival data can be categorised to be used as class labels or the survival data be regarded as another feature? It would be ideal if the survival data (and/or clinical data) can be incorporated into the clustering process to classify patients into subgroups or create subgroups that are both biologically meaningful and clinically relevant directly. So far, we found that latent supervised learning introduced by Wei and Kosorok [156] is able to do this. They proposed a binary classifier which uses a data-driven sieve maximum likelihood estimator for the separating hyperplane, which in turn can be used to estimate the parameters of the Gaussian mixture.

Investigation into an unsupervised feature selection technique based on the search for stable subgroups from different clustering algorithms (technical)

In Chapter 6, stable and clinically-useful breast cancer subgroups have been generated using a clustering algorithm on a reduced panel of 10 biomarkers [140]. When using ssFCM and the 15 features (obtained in Chapter 4.2 on all 1076 patients, Soria's classification [141] is fully retained and thus, biologically useful and clinically relevant. But the subgroups identified using ssFCM and the 15 features cannot be reproduced using unsupervised clus-

tering from CKM and MBIC. An investigation into an unsupervised feature selection technique based on stability of subgroups from different clustering algorithms as a feature selection criteria which will help answer the question if feature selection helps produce stable subgroups. Furthermore, it further ascertain whether the 10 features identified by Soria *et al.* [140] are the best 10 features for producing stable subgroups, or whether other features can be used to produce stable and clinically useful subgroups.

Application of integrated framework on other biomedical datasets (technical)

So far, the integrated framework was applied on the NTBC dataset despite the components within the framework being tested on several datasets. This is because configuration and investigation, that are specific to the dataset, are required (such as selection of suitable α setting and features selection). To demonstrate that the framework can solve real-world problems in other biomedical datasets, the application of the integrated framework on other biomedical datasets is considered for future work.

7.3 Dissemination of research

Accepted journal publications

1. D. T. C. Lai and J. M. Garibaldi. A Preliminary Study on Automatic Breast Cancer Data Classification using Semi-supervised Fuzzy c-Means. International Journal of Biomedical Engineering and Technology, Vol. 13, No. 4, pp. 303-322, 2013. (Chapter 3.2)

2. D. T. C. Lai, J. M. Garibaldi, D. Soria and C. M. Roadknight. A methodology for automatic classification of breast cancer immunohistochemical data using semi-supervised Fuzzy c-Means. Central European Journal of Operations Research, SI: Recent Advances in

Computational Biology, Bioinformatics, Medicine and Healthcare by Modern OR 2013 (in press). (Chapter 4.1 and 5)

Peer-reviewed and accepted conference papers

1. D. T. C. Lai and J. M. Garibaldi. A Comparison of Distance-based Semi-Supervised Fuzzy c-Means Clustering Algorithms. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Taipei, Taiwan, June 2011. pp. 1580 - 1586. (Chapter 3.1)

2. D. T. C. Lai and J. M. Garibaldi. Breast Cancer Data Classification using Semi-supervised Fuzzy c-means. Advances in Medical Signal and Information Processing (MEDSIP), Liverpool, UK, July 2012. (USB) (Chapter 3.2)

3. D. T. C. Lai and J. M. Garibaldi. Investigating Distance Metrics in Semi-supervised Fuzzy c-means for Breast Cancer Classification. Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB), Houston, USA, July 2012 - (LNCS 2013 Vol 7845, pp. 147-157). (Chapter 3.2)

4. H. Helmi, D. T. C. Lai and J. M. Garibaldi. Semi-Supervised Techniques in Breast Cancer Classification: A Comparison between Transductive SVM and Semi-Supervised FCM. 12th Annual Workshop on Computational Intelligence (UKCI), Edinburgh, UK, September 2012. (USB) (Chapter 3.3)

5. D. T. C. Lai and J. M. Garibaldi. Improving Semi-supervised Fuzzy C-Means Classification of Breast Cancer Data Using Feature Selection. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Hyderabad, India, July 2013. (Chapter 4.2)

6. D. T. C. Lai and J. M. Garibaldi. An investigation on scaling parameter and distance metrics in semi-supervised Fuzzy c-means. 12th Annual Workshop on Computational Intelligence (UKCI), Edinburgh, UK, September 2012. (USB) (Chapter 4.3)

7. D. T. C. Lai and J. M. Garibaldi. Identifying stable breast cancer subgroups using semi-supervised Fuzzy c-means on a reduced panel of biomarkers. 2014 IEEE World Congress on Computational Intelligence, Beijing, China, July 2014. (Chapter 6)

8. D. T. C. Lai and J. M. Garibaldi. Applying distance metric learning into a semi-supervised Fuzzy c-means framework. 2014 IEEE World Congress on Computational Intelligence, Beijing, China, July 2014. (Ongoing work on page 174)

Journal papers in preparation

1. D. T. C. Lai and J. M. Garibaldi. Approaches of improvement to semi-supervised Fuzzy c-means clustering with application on biomedical data. (Chapters 4, 6 and 7)

Presentations

Conference (oral) presentations:

1. Breast cancer classification using semi-supervised Fuzzy c-means, Advances in Medical Signal and Information Processing (MEDSIP), Liverpool, UK, July 2012.

2. Investigating Distance Metrics in Semi-supervised Fuzzy c-means for Breast Cancer Classification, Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB), Houston, USA, July 2012.

3. Fuzzy Covariance in Semi-supervised Fuzzy c-means, 2012 Mini EURO Conference on Computational Biology, Bioinformatics and Medicine (EURO-CBBM), Nottingham, UK, September 2012.

4. Improving Semi-supervised Fuzzy C-Means Classification of Breast Cancer Data Using Feature Selection, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Hyderabad, India, July 2013.

Intelligent Modelling and Analysis research group seminars:

1. Distance-based semi-supervised Fuzzy c-means algorithms, February 2011.

2. Breast cancer classification using semi-supervised Fuzzy c-means, February 2012.

3. Improving breast cancer classification using feature selection and semi-supervised Fuzzy c-means, November 2013.

Bibliography

[1] "NICE guidance recommends new test to guide breast cancer treatment decisions," 2013 Last assessed: 21st October 2013. [Online]. Available: http://www.nice.org.uk/newsroom/pressreleases/NICERecommendsTestBreastCancerTreatmentDecisions.jsp (Cited on pages 4 and 46.)

[2] "Association statistics," Last assessed: 7th May 2013. [Online]. Available: http://rss.acs.unt.edu/Rdoc/library/vcd/html/assocstats.html (Cited on pages xvi, 151, and 152.)

[3] 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," *International Journal of Cancer*, vol. 116, no. 3, pp. 340–350, 2005. (Cited on pages 1, 4, 5, 7, 46, 47, 48, 125, and 155.)

[4] S. P. Abney, *Semisupervised learning in computational linguistics*. CRC Press, 2008. (Cited on page 80.)

[5] A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964. (Cited on pages 25, 63, and 79.)

[6] M. B. Al-Daoud and S. A. Roberts, "New methods for the initialisation of clusters," *Pattern Recognition Letters*, vol. 17, pp. 451–455, May 1996. (Cited on page 39.)

[7] D. G. Altman and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity." *British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994. (Cited on page 146.)

[8] J. Augen, *Bioinformatics in the post-genomic era: Genome, transcriptome, proteome, and information-based medicine*. Addison-Wesley Professional, 2004. (Cited on pages 19 and 20.)

[9] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLoS Biology*, vol. 2, no. 4, p. e108, 2004. (Cited on pages 4, 17, 46, 119, and 142.)

[10] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005. (Cited on page 175.)

[11] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 27–34. (Cited on pages 57 and 60.)

[12] S. Basu, A. Banjeree, E. Mooney, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *SIAM International Conference on Data Mining*, 2004, pp. 333–344. (Cited on page 31.)

[13] O. Ben-Harush, I. Lapidot, and H. Guterman, "Initialization of iterative-based speaker diarization systems for telephone conversations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 414–425, 2012. (Cited on page 40.)

[14] M. Benkhalifa, A. Bensaid, and A. Mouradi, "Text categorization using the semi-supervised fuzzy c-means algorithm," in *International Conference of the North American Fuzzy Information Processing Society*, July 1999, pp. 561–565. (Cited on pages 40, 64, and 94.)

[15] A. M. Bensaid, L. O. Hall, J. C. Bezdek, and L. P. Clarke, "Partially supervised clustering for image segmentation," *Pattern Recognition*, vol. 29, no. 5, pp. 859–871, 1996. (Cited on pages 2, 28, 30, 34, 51, and 65.)

[16] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984. (Cited on pages 3 and 26.)

[17] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, 1981. (Cited on pages 14 and 21.)

[18] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines," *SIAM Journal on Applied Mathematics*, vol. 40, no. 2, pp. 339–357, 1981. (Cited on page 33.)

[19] J. Bezdek and N. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 28, no. 3, pp. 301–315, June 1998. (Cited on page 20.)

[20] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152. (Cited on page 79.)

[21] A. Bouchachia and W. Pedrycz, "A semi-supervised clustering algorithm for data exploration," *Fuzzy Sets and Systems - IFSA 2003 (International Fuzzy Systems Association World Congress Istanbul, Turkey)*, vol. 2715, pp. 107–155, 2003. (Cited on pages 2, 28, 30, 32, 33, 64, and 109.)

[22] ——, "Data clustering with partial supervision," *Data Mining and Knowledge Discovery*, vol. 12, pp. 47–78, 2006. (Cited on page 2.)

[23] ——, "Enhancement of fuzzy clustering by mechanisms of partial supervision," *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1733–1759, 2006. (Cited on pages 4, 16, 29, 30, 32, 34, 35, 51, 54, 59, 65, 109, and 110.)

[24] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional discriminant analysis," *Communications in Statistics - Theory and Methods*, vol. 36, no. 14, pp. 2607–2623, 2007. (Cited on page 79.)

[25] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering," in *International Conference on Machine Learning*, J. W. Shavlik, Ed. Morgan Kaufmann, 1998, pp. 91–99. (Cited on page 88.)

[26] F. Bray, J.-S. Ren, E. Masuyer, and J. Ferlay, "Global estimates of cancer prevalence for 27 sites in the adult population in 2008," *International Journal of Cancer*, vol. 132, no. 5, pp. 1133–1145, 2013. (Cited on page 4.)

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. (Cited on page 74.)

[28] N. Bundred, "Prognostic and predictive factors in breast cancer," *Cancer Treatment Reviews*, vol. 27, no. 3, pp. 137 – 142, 2001. (Cited on page 45.)

[29] M. Ceccarelli and A. Maratea, "Semi-supervised fuzzy c-means clustering of biological data," *Fuzzy Logic and Applications*, vol. 3849, pp. 259–266, 2006. (Cited on pages 28 and 35.)

[30] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-supervised learning*. Cambridge, MA, USA: MIT Press, 2006. (Cited on pages 2, 14, 26, and 80.)

[31] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," 2005. (Cited on page 80.)

[32] S. Chiu, "Fuzzy Model Identification based on cluster estimation," *Journal of Intelligent Fuzzy Systems*, vol. 2, pp. 267–278, 1994. (Cited on pages 38 and 88.)

[33] S. L. Chiu, *Fuzzy Information Engineering: A Guided Tour of Applications*. John Wiley and Sons, 1997, ch. Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification. (Cited on page 87.)

[34] M. Cianfrocca and L. J. Goldstein, "Prognostic and predictive factors in early-stage breast cancer," *The Oncologist*, vol. 9, no. 6, pp. 606–616, 2004. (Cited on page 45.)

[35] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, April 1960. (Cited on page 36.)

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: http://dx.doi.org/10.1007/BF00994018 (Cited on page 79.)

[37] H. Cramér, *Mathematical Methods of Statistics*. Princeton university press, 1999, vol. 9. (Cited on page 124.)

[38] P. Crewson, "The applied statistics handbook - coefficients for measuring association," 2012 Last assessed: 10th May 2014. [Online]. Available: http://www.acastat.com/Statbook/chisqassoc.htm (Cited on pages xv and 124.)

[39] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997. (Cited on page 41.)

[40] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979. (Cited on page 87.)

[41] G. Deboeck and T. Kohonen, *Visual explorations in finance: with self-organizing maps*. Springer Berlin, 1998, vol. 2. (Cited on page 17.)

[42] W. Du and K. Urahama, "Semi-supervised pattern classification utilizing fuzzy clustering and nonlinear mapping of data," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 11, pp. 1159–1164, 2007. (Cited on page 63.)

[43] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. John Wiley & Sons, 2001. (Cited on pages 14, 15, 18, 21, 24, 45, 72, and 140.)

[44] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973. (Cited on page 21.)

[45] ——, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974. (Cited on page 87.)

[46] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences USA*, no. 25, pp. 14 863–14 868, 1998. (Cited on pages 1, 17, and 46.)

[47] Y. Endo, Y. Hamasuna, M. Yamashiro, and S. Miyamoto, "On semi-supervised fuzzy c-means clustering," in *2009 IEEE International Conference on Fuzzy Systems*, 2009, pp. 1119–1124. (Cited on pages xiii, 2, 6, 28, 33, 34, 53, 54, 55, 56, 160, and 161.)

[48] M. Equihua, "Fuzzy clustering of ecological data," *Journal of Ecology*, vol. 78, no. 2, pp. 519–534, 1990. (Cited on page 17.)

[49] T. Etchells and P. J. G. Lisboa, "Orthogonal search-based rule extraction (osre) for trained neural networks: a practical and efficient approach," *IEEE Transactions on Neural Networks*, vol. 17, no. 2, pp. 374–384, 2006. (Cited on page 48.)

[50] B. Everitt, *Cluster analysis*, ser. Reviews of current research. Heinemann Educational [for] the Social Science Research Council, 1974. (Cited on pages 18 and 19.)

[51] J. Ferlay, H. Shin, F. Bray, D. Forman, C. Mathers, and D. Parkin, "Incidence/mortality data: Globocan 2008 v2.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 10 [internet], lyon, france: International agency for research on cancer; 2010." Last assessed: 21st June 2013. [Online]. Available: http://globocan.iarc.fr/factsheet.asp (Cited on page 4.)

[52] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936. (Cited on page 78.)

[53] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, no. 2, pp. 155–181, 2007. (Cited on pages 22 and 23.)

[54] A. Frank and A. Asuncion, "UCI machine learning repository [http://archive.ics.uci.edu/ml]," 2010. [Online]. Available: http://archive.ics.uci.edu/ml (Cited on pages 6, 53, and 58.)

[55] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, pp. 1–22, 2010. (Cited on page 76.)

[56] H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," *Pattern Recognition*, vol. 30, no. 7, pp. 1109–1119, 1997. (Cited on pages 2 and 31.)

[57] M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis, "The nottingham prognostic index in primary breast cancer," *Breast cancer research and treatment*, vol. 22, no. 3, pp. 207–219, 1992. (Cited on pages 45, 48, and 135.)

[58] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 148–155. (Cited on page 80.)

[59] C.-F. Gao and X.-J. Wu, "A new semi-supervised clustering algorithm with pairwise constraints by competitive agglomeration," *Appl. Soft Comput.*, vol. 11, no. 8, pp. 5281–5291, Dec. 2011. (Cited on pages 110, 161, and 166.)

[60] A. Globerson and S. Roweis, "Metric learning by collapsing classes," *Advances in Neural Information Processing Systems*, vol. 18, pp. 451–458, 2006. (Cited on page 174.)

[61] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," *Advances in Neural Information Processing Systems*, pp. 513–520, 2004. (Cited on page 174.)

[62] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83–90, 2006. (Cited on page 41.)

[63] A. Green, D. Powe, E. Rakha, D. Soria, C. Lemetre, C. Nolan, F. Barros, R. Macmillan, J. Garibaldi, G. Ball, and I. Ellis, "Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers," *British Journal of Cancer*, 2013. (Cited on pages iv, 9, 50, 141, 142, 146, 155, 157, 158, and 170.)

[64] N. Grira, M. Crucianu, and N. Boujemaa, "Semi-supervised fuzzy clustering with pairwise-constrained competitive agglomeration," in *2005 IEEE International Conference on Fuzzy Systems*, May 2005, pp. 867–872. (Cited on pages 2, 31, 34, 161, and 172.)

[65] ——, "Active semi-supervised fuzzy clustering," *Pattern Recognition.*, vol. 41, pp. 1851–1861, May 2008. (Cited on pages 31 and 65.)

[66] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, vol. 17, January 1978, pp. 761–766. (Cited on pages 24, 25, and 27.)

[67] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003. (Cited on pages 41, 44, and 108.)

[68] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002. (Cited on pages 40, 41, 43, and 44.)

[69] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate data analysis: A global perspective*, 7th ed. Pearson Education, 2010. (Cited on pages 35 and 136.)

[70] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001. (Cited on pages 3 and 15.)

[71] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999. (Cited on pages 40, 41, and 107.)

[72] M. A. Hall and L. A. Smith, "Feature subset selection: a correlation based filter approach," in *1997 International Conference on Neural Information Processing and Intelligent Information Systems*, 1997, pp. 855–858. (Cited on page 41.)

[73] D. P. Harrington and T. R. Fleming, "A class of rank test procedures for censored survival data," *Biometrika*, vol. 69, no. 3, pp. 553–566, 1982. (Cited on page 155.)

[74] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 155–176, 1996. (Cited on page 78.)

[75] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The elements of statistical learning*. Springer New York, 2001, vol. 1. (Cited on page 4.)

[76] J. Haybittle, R. Blamey, C. Elston, J. Johnson, P. Doyle, F. Campbell, R. Nicholson, and K. Griffiths, "A prognostic index in primary breast cancer." *British journal of cancer*, vol. 45, no. 3, p. 361, 1982. (Cited on page 45.)

[77] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Prentice Hall, 1998. (Cited on page 48.)

[78] J. He, M. Lan, C. L. Tan, S. Y. Sung, and H. B. Low, "Initialization of cluster refinement algorithms: a review and comparative study," in *IEEE International Joint Conference on Neural Networks*, vol. 1, July 2004, pp. 297–302. (Cited on pages 38, 39, 40, and 88.)

[79] H. Helmi, D. T. C. Lai, and J. M. Garibaldi, "Semi-supervised techniques in breast cancer classification: A comparison between tranductive SVM and semi-supervised FCM," in *12th Annual Workshop on Computational Intelligence (UKCI)*, 2012. (Cited on pages xiii, 81, and 82.)

[80] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. (Cited on page 76.)

[81] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Wiley, 1999. (Cited on pages 22, 26, and 32.)

[82] P. Huang and D. Zhang, "Locality sensitive c-means clustering algorithms," *Neurocomputing*, vol. 73, no. 16-18, pp. 2935–2943, 2010, 10th Brazilian Symposium on Neural Networks (SBRN2008). (Cited on page 31.)

[83] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, pp. 264–323, September 1999. (Cited on page 38.)

[84] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010. (Cited on pages 2, 15, 16, 26, 34, 140, 162, 168, and 172.)

[85] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000. (Cited on pages 4, 24, 26, 64, and 162.)

[86] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in kernel methods*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 169–184. (Cited on page 80.)

[87] ——, "Transductive inference for text classification using support vector machines," in *International Conference on Machine Learning*, vol. 99, 1999, pp. 200–209. (Cited on page 79.)

[88] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345. (Cited on page 48.)

[89] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007. (Cited on pages 44 and 45.)

[90] I. Katsavounidis, C.-C. Jay Kuo, and Z. Zhang, "A new initialization technique for generalized lloyd iteration," *Signal Processing Letters, IEEE*, vol. 1, no. 10, pp. 144–146, October 1994. (Cited on pages 9 and 39.)

[91] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, 2005. (Cited on page 14.)

[92] T. Kohonen, *Self-organizing maps*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997. (Cited on page 77.)

[93] ——, *Learning Vector Quantization*, ser. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2001, vol. 30. (Cited on page 77.)

[94] Y. Q. Kong and S. T. Wang, "Feature selection and semisupervised fuzzy clustering," *Fuzzy Information and Engineering*, vol. 1, pp. 179–190, 2009. (Cited on page 28.)

[95] M. Kuhn, "Variable selection using the `caret` package," 2011 Last assessed: 5th November 2013. [Online]. Available: http://cran.r-project.org/web/packages/caret/vignettes/caretSelection.pdf (Cited on pages 41, 43, 44, 108, and 146.)

[96] ——, "The `caret` package," 2009 Last assessed: 5th November 2013. [Online]. Available: http://cran.r-project.org/web/packages/caret/caret.pdf (Cited on page 80.)

[97] D. T. C. Lai and J. M. Garibaldi, "A comparison of distance-based semi-supervised fuzzy c-means clustering algorithms," in *2011 IEEE International Conference on Fuzzy Systems*, June 2011, pp. 1580–1586. (Cited on page 59.)

[98] ——, "Improving semi-supervised fuzzy c-means classification of breast cancer data using feature selection," in *2013 IEEE International Conference on Fuzzy Systems*, 2013, pp. 1–8. (Cited on pages 121 and 141.)

[99] ——, "A preliminary study on automatic breast cancer data classification using semi-supervised fuzzy c-means," *International Journal of Biomedical Engineering and Technology*, vol. 13, no. 4, pp. 303–322, 2013. (Cited on pages 65, 72, 143, and 162.)

[100] D. T. C. Lai, J. M. Garibaldi, D. Soria, and C. M. Roadknight, "A methodology for automatic classification of breast cancer immunohistochemical data using semi-supervised fuzzy c-means," *Central European Journal of Operations Research*, pp. 1–25, 2013. (Cited on pages 72, 121, and 138.)

[101] C. Li, L. Liu, and W. Jiang, "Objective function of semi-supervised fuzzy c-means clustering algorithm," in *IEEE International Conference on Industrial Informatics*, July 2008, pp. 737–742. (Cited on pages xiii, 6, 16, 28, 29, 34, 35, 40, 51, 53, 54, 55, and 160.)

[102] K. Li, Z. Cao, L. Cao, and R. Zhao, "A novel semi-supervised fuzzy c-means clustering method," in *Proceedings of Chinese Control and Decision Conference, CCDC '09.*, June 2009, pp. 3761 –3765. (Cited on page 57.)

[103] H. Liu and S. T. Huang, "Evolutionary semi-supervised fuzzy clustering," *Pattern Recognition Letters*, vol. 24, pp. 3105–3113, December 2003. (Cited on page 2.)

[104] W. Y. Liu, C. J. Xiao, B. W. Wang, Y. Shi, and S. F. Fang, "Study on combining subtractive clustering with fuzzy c-means clustering," in *International Conference on Machine Learning and Cybernetics 2003*, vol. 5, Nov 2003, pp. 2659–2662. (Cited on page 88.)

[105] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003. (Cited on pages 17, 19, and 34.)

[106] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, University of California Press, 1967, pp. 281–297. (Cited on page 19.)

[107] R. D. Maesschalck, D. Jouan-Rimbaud, and D. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000. (Cited on page 24.)

[108] N. Makretsov, D. Huntsman, T. Nielsen, E. Yorida, M. Peacock, M. Cheang, S. Dunn, M. Hayes, M. van de Rijn, C. Bajdik, and C. Gilks, "Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma," *Clinical Cancer Research*, vol. 10, no. 18, pp. 6143–6151, 2004. (Cited on page 46.)

[109] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. (Cited on page 35.)

[110] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, December 2002. (Cited on page 20.)

[111] W. L. McGuire, "Breast cancer prognostic factors: evaluation guidelines," *Journal of the National Cancer Institute*, vol. 83, no. 3, pp. 154–155, 1991. (Cited on page 45.)

[112] D. Meyer, A. Zeileis, K. Hornik, F. Gerber, and M. Friendly, "The `vcd` package," 2013 Last assessed: 13th December 2013. [Online]. Available: http://cran.r-project.org/web/packages/vcd/vcd.pdf (Cited on page 124.)

[113] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005. (Cited on page 48.)

[114] T. M. Mitchell, *Machine learning*. McGraw-Hill Boston, MA, 1997. (Cited on page 82.)

[115] A. Y. Ng, "Preventing "overfitting" of cross-validation data," in *International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 245–253. (Cited on pages 44 and 98.)

[116] R. Nicholson, J. Gee, M. . Harper *et al.*, "EGFR and cancer prognosis." *European Journal of Cancer (Oxford, England: 1990)*, vol. 37, p. S9, 2001. (Cited on page 108.)

[117] N. Päivinen, "Clustering with a minimum spanning tree of scale-free-like structure," *Pattern Recognition Letters.*, vol. 26, pp. 921–930, May 2005. (Cited on page 64.)

[118] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, 2005. (Cited on page 40.)

[119] N. Pal and J. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, Aug 1995. (Cited on page 20.)

[120] J.-M. Park and H.-D. Yae, "Analysis of active feature selection in optic nerve data using labeled fuzzy c-means clustering," in *2002 IEEE International Conference on Fuzzy Systems*, vol. 2. IEEE, 2002, pp. 1580–1585. (Cited on pages 40 and 94.)

[121] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 27, no. 5, pp. 787–795, May 1997. (Cited on pages vi, xiii, 2, 6, 16, 18, 25, 27, 28, 29, 30, 32, 34, 35, 51, 53, 54, 55, 59, 60, 65, 70, 88, 110, 122, 160, 161, 172, and 173.)

[122] W. Pedrycz, "Algorithms of fuzzy clustering with partial supervision," *Pattern Recognition Letters*, vol. 3, no. 1, pp. 13–20, 1985. (Cited on pages 27, 28, 29, 32, and 53.)

[123] ——, "Fuzzy sets in pattern recognition: Methodology and methods," *Pattern Recognition*, vol. 23, no. 1-2, pp. 121–146, 1990. (Cited on page 2.)

[124] C. Perou, T. Sørlie, M. Eisen, M. Van de Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, H. Johnsen, L. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. Zhu, P. Lønning, A. Børresen-Dale, P. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, 2000. (Cited on page 46.)

[125] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009. (Cited on page 74.)

[126] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. (Cited on pages 48 and 74.)

[127] R. Quinlan, "Is see5/c5.0 better than c4.5?" 2012 Last assessed: 29 July 2013. [Online]. Available: http://www.rulequest.com/see5-comparison.html (Cited on page 74.)

[128] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org (Cited on page 58.)

[129] R Core Team and contributors worldwide, "Correlation, variance and covariance (matrices)," 2012 Last assessed: 12th May 2014. [Online]. Available: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.html (Cited on page 141.)

[130] ——, "Kruskal-wallis rank sum test," 2012 Last assessed: 12th May 2014. [Online]. Available: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/kruskal.test.html (Cited on pages xviii and 201.)

[131] ——, "Wilcoxon rank sum and signed rank tests," 2013 Last assessed: 28th November 2013. [Online]. Available: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html (Cited on page 114.)

[132] E. A. Rakha, D. Soria, C. Lemetre, A. R. Green, D. G. Powe, C. C. Nolan, J. M. Garibaldi, G. R. Ball, and I. O. Ellis, "Nottingham prognostic index plus (npi+): a modern clinical decision making tool in breast cancer," *British Journal of Cancer (In submission)*, 2013. (Cited on pages xiv, 51, 102, 104, and 142.)

[133] B. D. Ripley, *Pattern recognition and neural networks*. New York: Cambridge university press, 2007. (Cited on page 77.)

[134] ——, "nnet: Feed-forward neural networks and multinomial log-linear models," 2013 Last assessed: 26th November 2013. [Online]. Available: http://cran.r-project.org/web/packages/nnet/nnet.pdf (Cited on page 77.)

[135] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. (Cited on page 41.)

[136] S. J. Schnitt, "Traditional and newer pathologic factors," *Journal of the National Cancer Institute Monographs*, vol. 2001, no. 30, pp. 22–26, 2001. (Cited on page 45.)

[137] B. Schölkopf, "The kernel trick for distances," *Advances in Neural Information Processing Systems*, pp. 301–307, 2001. (Cited on pages 25 and 63.)

[138] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978. (Cited on page 23.)

[139] I. Soerjomataram, J. Lortet-Tieulent, D. M. Parkin, J. Ferlay, C. Mathers, D. Forman, and F. Bray, "Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions," *The Lancet*, 2012. (Cited on page 4.)

[140] D. Soria, J. Garibaldi, E. Biganzoli, and I. Ellis, "A comparison of three different methods for classification of breast cancer data," in *Seventh International Conference on Machine Learning and Applications*, 2008, pp. 619–624. (Cited on pages 9, 35, 48, 96, 141, 142, 145, 147, 148, 157, 171, 176, and 177.)

[141] D. Soria, J. M. Garibaldi, F. Ambrogi, A. R. Green, D. Powe, E. Rakha, R. D. Macmillan, R. W. Blamey, G. Ball, P. J. 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," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 318–330, 2010. (Cited on pages iv, xi, xii, xiii, xv, xvi, xvii, 1, 4, 5, 6, 7, 10, 11, 13, 34, 35, 45, 47, 48, 50, 51, 52, 54, 66, 67, 68, 70, 71, 72, 95, 108, 120, 123, 124, 125, 126, 127, 128, 129, 132, 135, 137, 139, 142, 145, 147, 150, 152, 154, 156, 157, 159, 162, 165, 167, 168, 169, 176, 199, and 203.)

[142] D. Soria, J. M. Garibaldi, A. R. Green, D. G. Powe, C. C. Nolan, C. Lemetre, G. R. Ball, and I. O. Ellis, "A quantifier-based fuzzy classification system for breast cancer patients," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 175–184, 2013. (Cited on pages 50 and 141.)

[143] T. Sørlie, C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. Eisen, M. van de Rijn, S. Jeffrey, T. Thorsen, H. Quist, J. Matese, P. Brown, D. Botstein, P. Eystein Lønning, and A. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences USA*, vol. 98, no. 19, pp. 10 869–10 874, 2001. (Cited on page 46.)

[144] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geister, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A. Børresen-Dale, and D. Botstein, "Repeated observation of breast tumor subtypes in indepedent gene expression data sets," *Proceedings of the National Academy of Sciences USA*, vol. 100, no. 14, pp. 8418–8423, 2003. (Cited on page 46.)

[145] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, March 2003. (Cited on page 36.)

[146] C. Stutz and T. A. Runkler, "Classification and prediction of road traffic using application-specific fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 3, pp. 297–308, June 2002. (Cited on pages 17, 29, 32, 34, and 35.)

[147] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining.* Pearson Education India, 2007. (Cited on pages 16 and 20.)

[148] L. Tari, C. Baral, and S. Kim, "Fuzzy c-means clustering with prior biological knowledge," *Journal of Biomedical Informatics*, vol. 42, no. 1, pp. 74–81, 2009. (Cited on pages 34, 35, and 65.)

[149] R. Taylor, "Interpretation of the correlation coefficient: a basic review," *Journal of Diagnostic Medical Sonography*, vol. 6, no. 1, pp. 35–39, 1990. (Cited on page 141.)

[150] T. Therneau, "The `survival` package," 2013 Last assessed: 13th December 2013. [Online]. Available: http://cran.r-project.org/web/packages/survival/survival.pdf (Cited on pages xvii, xviii, 155, 200, 201, 202, 203, and 204.)

[151] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996. (Cited on page 76.)

[152] L. Toloşi and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011. (Cited on page 108.)

[153] J. Tou and R. Gonzales, *Pattern Recognition Principles.* Addison-Wesley, Reading, MA, 1974. (Cited on page 39.)

[154] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002. (Cited on pages 1, 4, and 46.)

[155] N. Wang, X. Li, and X. Luo, "Semi-supervised kernel-based fuzzy c-means with pairwise constraints," in *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, June 2008, pp. 1098–1102. (Cited on pages 28, 31, 110, and 166.)

[156] S. Wei and M. R. Kosorok, "Latent supervised learning," *Journal of the American Statistical Association*, vol. 108, no. 503, pp. 957–970, 2013. (Cited on page 176.)

[157] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Advances in Neural Information Processing Systems*, pp. 1473–1480, 2006. (Cited on page 174.)

[158] M. P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets and Systems*, vol. 5, no. 2, pp. 177–185, 1981. (Cited on page 20.)

[159] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 3rd ed. Burlington, Massachusetts, USA: Morgan Kaufmann, 2011. (Cited on page 18.)

[160] J. Wu, H. Xiong, C. Liu, and J. Chen, "A generalization of distance functions for fuzzy c-means clustering with centroids of arithmetic means," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 3, pp. 557–571, 2012. (Cited on page 175.)

[161] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions Pattern Analysis and Machine Intelligence.*, vol. 13, pp. 841–847, August 1991. (Cited on page 3.)

[162] Z. Xie, Q. Hu, and D. Yu, "Improved feature selection algorithm based on svm and correlation," *Advances in Neural Networks - ISNN 2006*, pp. 1373–1380, 2006. (Cited on page 107.)

[163] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," *Advances in Neural Information Processing Systems*, vol. 15, pp. 505–512, 2002. (Cited on pages 24 and 174.)

[164] R. Yager and D. Filev, "Approximate clustering via the mountain method," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 8, pp. 1279–1284, August 1994. (Cited on page 87.)

[165] D. Zhang, K. Tan, and S. Chen, "Semi-supervised kernel-based fuzzy c-means," *Lecture Notes in Computer Science: Neural Information Processing*, vol. 3316, pp. 1229–1234, July 2004. (Cited on pages xiii, 2, 6, 26, 28, 31, 32, 53, 54, 55, 58, 160, and 161.)

[166] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009. (Cited on pages 2, 14, 26, and 27.)

Appendix

Table 7.1: Classification results from running four ssFCM algorithms. The ssFCM with the highest result for each dataset is indicated in bold.

	% labelled					
dataset	2	4	6	8	10	20
Results produced by Pedrycz-97						
Iris-4	77.33	81.33	82.00	86.00	84.00	90.00
Iris-2	62.00	75.33	74.00	77.33	79.33	85.33
Wine-13	55.06	65.17	71.91	71.91	78.65	84.27
Wine-2	**80.34**	**80.90**	**81.46**	80.90	81.46	82.58
XOR-2	**56.00**	74.50	84.00	93.00	86.50	91.50
WOBC-8	79.54	84.12	87.84	88.70	87.55	90.99
WOBC-2	89.84	**94.99**	**94.85**	**94.56**	**94.85**	**96.14**
PID-8	54.17	**60.03**	**66.41**	**65.49**	**70.05**	**75.91**
PID-2	51.30	53.26	54.17	54.43	54.82	57.55
WDBC-30	80.67	73.64	81.90	83.48	88.22	91.56
Results produced by Li-08						
Iris-4	**94.67**	**95.33**	**98.00**	**98.00**	**98.00**	**98.00**
Iris-2	64.67	**94.00**	**94.00**	**94.00**	**94.00**	**96.67**
Wine-13	60.11	61.80	66.85	**87.08**	**89.89**	89.33
Wine-2	68.54	**80.90**	**81.46**	**83.15**	**82.58**	**86.52**
XOR-2	50.00	**99.50**	**99.50**	**99.50**	**99.50**	**99.50**
WOBC-8	**88.27**	88.70	80.54	80.97	81.55	89.99
WOBC-2	**92.85**	93.56	93.56	93.71	93.99	95.42
PID-8	43.23	47.92	50.39	53.78	56.51	70.31
PID-2	44.14	47.27	48.57	50.91	52.34	61.20
WDBC-30	85.59	**92.79**	**90.86**	**90.69**	**89.81**	**93.67**
Results produced by Zhang-04						
Iris-4	66.67	72.00	72.00	74.00	75.33	85.33
Iris-2	**68.00**	72.67	73.33	76.00	76.67	84.67
Wine-13	55.62	56.18	56.18	57.30	74.16	76.40
Wine-2	29.21	75.28	76.40	81.46	80.90	84.27
XOR-2	48.50	60.50	54.00	60.50	63.00	66.50
WOBC-8	87.55	**89.41**	**90.27**	**91.56**	**92.13**	**96.14**
WOBC-2	92.56	92.56	92.70	93.13	93.42	94.13
PID-8	**55.60**	57.16	57.55	58.20	59.38	64.71
PID-2	56.12	57.81	58.72	58.98	60.29	64.84
WDBC-30	**87.17**	87.87	88.40	88.58	88.75	89.46
Results produced by Endo-09						
Iris-4	67.33	69.33	74.00	74.00	74.00	78.67
Iris-2	66.67	68.67	70.00	72.67	74.00	76.67
Wine-10[1]	**87.64**	**89.89**	**91.01**	86.52	89.33	**91.57**
Wine-2	50.56	71.35	79.21	82.58	82.02	83.71
XOR-2	26.50	28.00	29.50	31.00	32.50	40.00
WOBC-8	66.24	67.53	68.38	69.81	70.82	76.54
WOBC-2	66.24	67.53	68.38	69.81	70.82	76.54
PID-6 [1]				*		
PID-2	**65.36**	**66.02**	**66.93**	**67.45**	**68.62**	**71.61**
WDBC-23 [1]	63.27	64.50	65.73	66.96	68.01	71.88

[1] Features have to be reduced due to algorithm failure caused by infinity problem

[*] Algorithm failure

Table 7.2: Accuracy of ssFCM using Euclidean (E), Mahalanobis(M), Fuzzy Mahalanobis (FM) and kernel-based (K) distances based on CV. The distance metric with highest average accuracy, κ and NMI is indicated in italics, showing that Euclidean with ssFCM is most suitable for NTBC.

DM[1]	ET[2]	0%	10%	20%	30%	40%	50%	60%
E	A[3]	30.47±0.46	*96.12±2.04*	*96.86±1.94*	*97.22±1.77*	*97.54±1.61*	*97.64±1.55*	*97.84±1.53*
	κ[4]	0.00±0.00	*0.95±0.03*	*0.96±0.02*	*0.97±0.02*	*0.97±0.02*	*0.97±0.02*	*0.97±0.02*
	NMI[5]	NaN	*0.93±0.04*	*0.94±0.03*	*0.95±0.03*	*0.95±0.03*	*0.96±0.03*	*0.96±0.03*
M	A	30.47±0.46	75.00±5.66	81.39±5.64	84.61±5.08	85.95±5.11	86.88±4.88	87.63±4.86
	κ	0.00±0.00	0.69±0.07	0.77±0.07	0.81±0.06	0.83±0.06	0.84±0.06	0.85±0.06
	NMI	NaN	0.61±0.08	0.69±0.08	0.73±0.08	0.75±0.08	0.77±0.08	0.78±0.08
FM	A	30.47±0.46	35.12±7.41	38.89±6.45	41.93±5.26	43.71±4.91	46.09±4.83	48.02±5.46
	κ	0.00±0.00	0.18±0.07	0.22±0.07	0.25±0.06	0.27±0.06	0.30±0.06	0.32±0.07
	NMI	NaN	0.34±0.06	0.36±0.06	0.40±0.06	0.38±0.06	0.41±0.06	0.43±0.06
K	A	30.47±0.46	69.59±6.20	76.46±5.12	81.35±5.09	82.14±4.99	82.28±5.25	83.47±4.85
	κ	0.00±0.00	0.60±0.08	0.69±0.07	0.76±0.07	0.77±0.07	0.77±0.07	0.79±0.06
	NMI	NaN	0.55±0.08	0.63±0.07	0.69±0.08	0.70±0.07	0.70±0.08	0.72±0.07

[1] Distance Metric
[2] Evaluation Technique
[3] Accuracy in percentage
[4] Cohen's Kappa Index
[5] Normalised Mutual Index

Table 7.3: Average training accuracy of ssFCM on NTBC using all features and 15 selected features

	10%	20%	30%	40%	50%	60%
All features	96.71±1.03	97.93±0.66	98.52±0.49	98.95±0.41	99.23±0.30	99.48±0.24
SVM-RFE-15	95.68±2.17	97.19±1.03	97.92±0.64	98.29±0.65	98.73±0.41	99.09±0.32
CFS-15	96.49±1.72	97.21±0.98	97.97±0.80	98.46±0.73	98.79±0.64	99.14±0.44
NB-RFE-15	96.73±1.33	97.86±0.59	98.39±0.45	98.81±0.38	99.07±0.31	99.30±0.26
RF-RFE-15	96.57±1.17	97.73±0.68	98.28±0.56	98.75±0.49	99.00±0.37	99.31±0.30
IG-15	93.55±4.35	97.23±1.38	98.05±0.64	98.54±0.50	98.88±0.36	99.21±0.29
GR-15	93.79±4.33	97.07±1.30	97.65±1.17	98.29±0.69	98.55±0.88	99.10±0.39
CSQ-15	93.59±4.39	97.19±1.15	97.99±0.71	98.53±0.51	98.88±0.37	99.22±0.30

Table 7.4: Rank count of 15 selected features from NB-RFE which achieved 100% accuracy with ssFCM on NTBC.

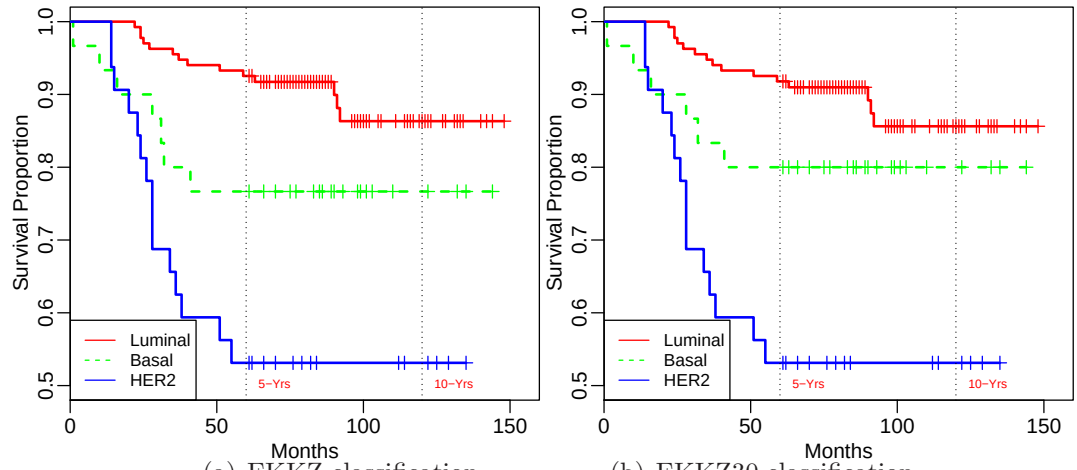
	rank																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Score	Rank
CK18	104	71	13	2	1	1	0	0	0	0	0	0	0	0	0	2768	1
CK7/8	73	94	16	7	2	0	0	0	0	0	0	0	0	0	0	2725	2
p53	6	14	122	27	9	8	4	1	0	1	0	0	0	0	0	2425	3
MUC1co	2	4	16	74	59	21	9	6	1	0	0	0	0	0	0	2177	4
ER	1	7	12	42	52	34	27	14	3	0	0	0	0	0	0	2061	5
MUC1	0	0	0	7	31	64	52	28	8	1	0	1	0	0	0	1823	6
HER2	6	2	8	19	20	30	34	39	22	5	3	2	1	1	0	1800	7
PgR	0	0	3	8	13	23	46	63	33	2	1	0	0	0	0	1674	8
CK19	0	0	2	5	5	9	16	35	104	15	0	0	1	0	0	1476	9
HER4	0	0	0	1	0	1	3	4	15	149	14	5	0	0	0	1170	10
AR	0	0	0	0	0	0	0	1	4	7	78	88	13	1	0	861	11
HER3	0	0	0	0	0	1	1	1	2	10	84	62	24	4	2	851	12
nBRCA1	0	0	0	0	0	0	0	0	0	1	11	29	116	26	9	586	13
FHIT	0	0	0	0	0	0	0	0	0	1	1	4	20	89	69	334	14
CK5/6	0	0	0	0	0	0	0	0	0	0	0	1	17	68	91	282	15
E-cad	0	0	0	0	0	0	0	0	0	0	0	0	0	2	14	18	16
EGFR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	17
Actin	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	17
GCDFP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	19
P-cad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	19
CK14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	19
MUC2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22
p63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22
Synapto	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22
Chromo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22

Table 7.5: Accuracy using Euclidean ssFCM and α values on NTBC in a clustering setting.

α	10%	20%	30%	40%	50%	60%
N/M	**96.51±1.32**	97.74±0.68	98.41±0.53	98.73±0.43	99.05±0.45	99.24±0.29
0.1	90.00±4.85	94.74±1.07	96.48±0.59	97.16±0.39	97.69±0.39	98.12±0.35
1	95.05±2.93	97.19±0.66	98.10±0.58	98.54±0.45	98.94±0.46	99.19±0.30
10	**96.51±1.33**	**97.83±0.67**	**98.52±0.52**	**98.83±0.43**	**99.13±0.41**	**99.31±0.28**

Table 7.6: Class distributions of patients used in survival analysis based on methodology by Soria *et al.* [141], EKKZ, ENB, EKKZ30 and ENB30.

	1	2	3	4	5	6	total
Soria	71(202)	90(153)	43(80)	36(82)	38(69)	44(77)	322 (663)
EKKZ	39(91)	68(113)	27(66)	5(15)	25(58)	32(70)	199 (413)
ENB	46(95)	60(108)	27(64)	5(15)	25(58)	33(73)	199 (413)
EKKZ30	48(107)	66(111)	21(56)	5(13)	25(59)	31(67)	199 (413)
ENB30	49(109)	65(110)	21(55)	5(13)	25(59)	31(67)	199 (413)



(a) EKKZ classification (b) EKKZ30 classification

(c) ENB classification (d) ENB30 classification

Figure 7.1: Kaplan-Meier analysis of overall survival for classifying 413 patients using various ssFCM methodologies, showing 3 main groups where the three survival curves are visibly well-separated.

Table 7.7: Survival curve differences using log-rank test [150] based on Soria's classification with 6 subgroups and 3 main groups. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

Soria's classification 6 subgroups *						3 main groups *		
Class	1	2	3	4	5	Class	1	2
2	0.978					2	*	
3	0.365	0.519				3	*	0.217
4	*	*	0.001					
5	*	*	0.001	0.858				
6	*	*	*	0.345	0.259			

* p< 0.001

Table 7.8: Survival curve differences using log-rank test [150] based on EKKZ classification of 413 patients. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

EKKZ classification 6 subgroups *						3 main groups *		
Class	1	2	3	4	5	Class	1	2
2	0.515					2	0.040	
3	0.594	0.99				3	*	0.067
4	0.426	0.663	0.533					
5	0.055	0.102	0.199	0.853				
6	*	*	0.002	0.286	0.100			

* p< 0.001

Table 7.9: Survival curve differences using log-rank test [150] based on EKKZ30 classification of 413 patients. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

EKKZ30 classification 6 subgroups *						3 main groups *		
Class	1	2	3	4	5	Class	1	2
2	0.966					2	0.154	
3	0.486	0.451				3	*	0.035
4	0.545	0.591	0.736					
5	0.217	0.162	0.623	0.986				
6	*	*	0.009	0.286	0.051			

* p< 0.001

Table 7.10: Survival curve differences using log-rank test [150] based on ENB classification of 413 patients. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

ENB classification 6 subgroups *						3 main groups *		
Class	1	2	3	4	5	Class	1	2
2	0.863					2	0.158	
3	0.609	0.479				3	*	0.043
4	0.599	0.577	0.709					
5	0.25	0.157	0.581	0.986				
6	*	*	0.009	0.307	0.062			

* p< 0.001

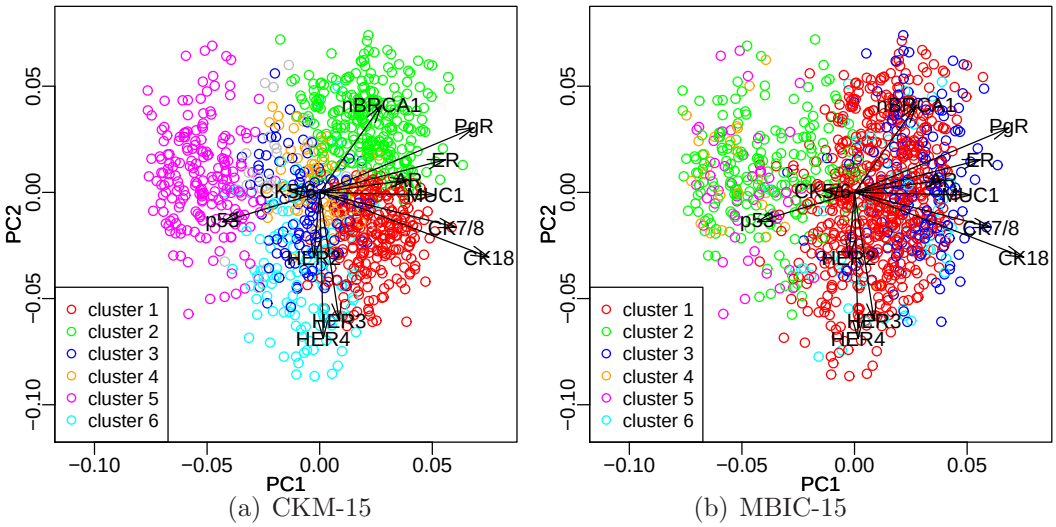
Table 7.11: Survival curve differences using log-rank test [150] based on ENB30 classification of 413 patients. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

ENB30 classification 6 subgroups *						3 main groups *		
Class	1	2	3	4	5	Class	1	2
2	0.968					2	0.189	
3	0.283	0.276				3	*	0.046
4	0.542	0.612	0.042					
5	0.203	0.185	0.882	0.986				
6	*	*	0.042	0.307	0.065			

* p< 0.001

Table 7.12: Comparison of NPI distribution between Soria's classification and ssFCMs classification using Kruskal-Wallis test [130] where p>= 0.01 accepts the null hypothesis, indicating that the two populations have identical distribution.

	c1	c2	c3	c4	c5	c6
EKKZ	0.125	0.024	0.826	0.693	0.559	0.980
ENB	0.197	0.010	0.661	0.693	0.656	0.784
EKKZ30	0.031	0.045	0.755	0.249	0.561	0.894
ENB30	0.037	0.034	0.922	0.249	0.561	0.860



(a) CKM-15 (b) MBIC-15

Figure 7.2: Biplots based on clustering 1076 patients using the 15 important features with CKM in (a) and with MBIC in (b).

Table 7.13: Correlation between protein biomarkers and clinical data.

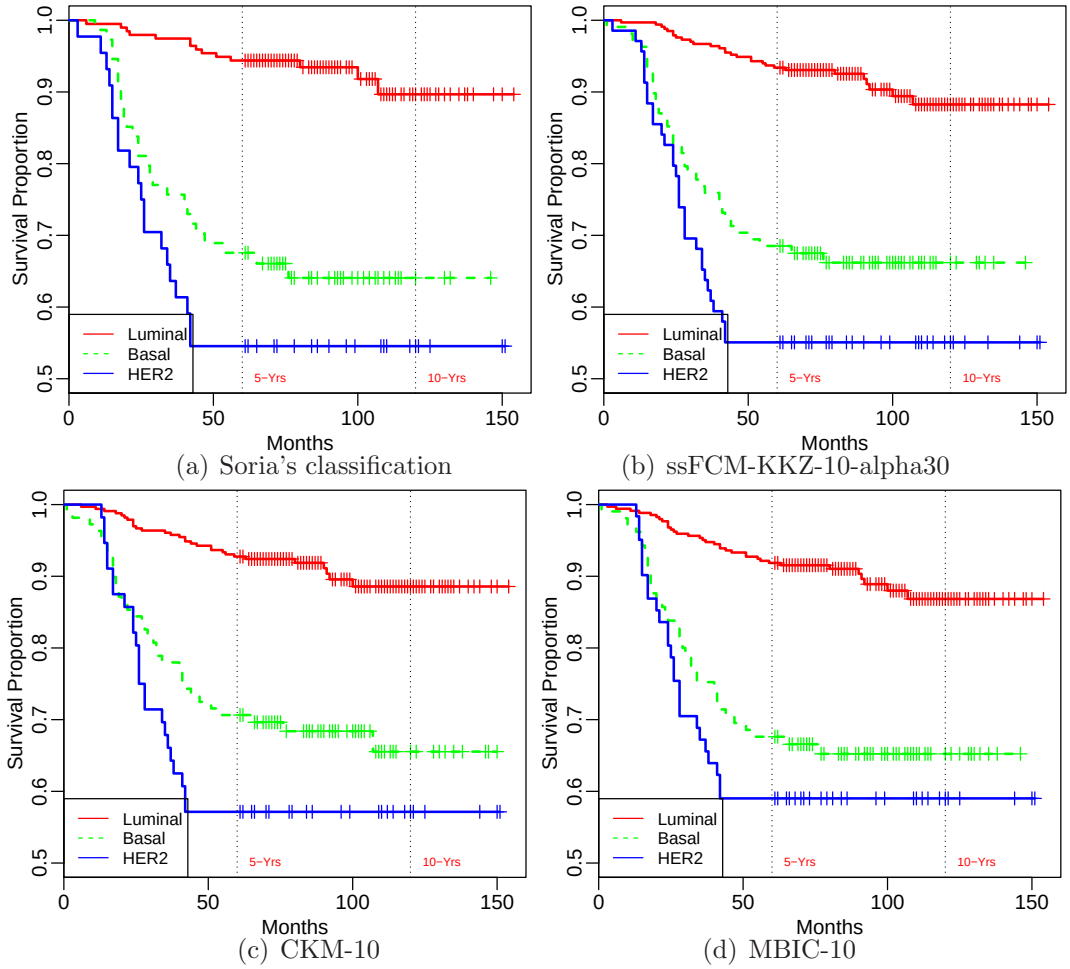
Biomarker/ Clinical data	Age	Grade	Size	Stage	NPI	Death
CK7/8	*	-0.376	*	*	*	*
CK18	*	-0.338	*	*	*	*
CK19	*	-0.340	*	*	*	*
CK5/6	*	*	*	*	*	*
CK14	*	*	*	*	*	*
Actin	*	*	*	*	*	*
p63	*	*	*	*	*	*
ER	0.340	-0.338	*	*	*	*
PgR	*	-0.394	*	*	-0.300	*
AR	*	*	*	*	*	*
EGFR	*	*	*	*	*	*
HER2	*	*	*	*	*	*
HER3	*	*	*	*	*	*
HER4	*	*	*	*	*	*
p53	*	0.365	*	*	*	*
nBRCA1	*	*	*	*	*	*
FHIT	*	*	*	*	*	*
E-cad	*	*	*	*	*	*
P-cad	*	*	*	*	*	*
MUC1	*	*	*	*	*	*
MUC1co	*	-0.343	*	*	*	*
MUC2d	*	*	*	*	*	*
GCDFP	*	*	*	*	*	*
Chromo	*	*	*	*	*	*
Synapto	*	*	*	*	*	*

* $-0.3 <$coef< 0.3 - very weakly related

Table 7.14: Survival curve differences using log-rank test [150] based on Soria's classification. The first line shows p-value based on comparison with all survival curves together using the two different grouping, 7 subgroups and 3 main groups.

CKM-10 7 subgroups *							3 main groups *		
Cluster	1	2	3	4	5	6	Cluster	1	2
2	0.978						2	*	
3	0.365	0.519					3	*	0.217
4	*	*	0.001						
5	*	*	0.001	0.858					
6	0.001	*	*	0.516	0.398				
7	*	*	*	0.377	0.298	0.897			

* $p< 0.001$



(a) Soria's classification (b) ssFCM-KKZ-10-alpha30

(c) CKM-10 (d) MBIC-10

Figure 7.3: Survival curves based on 3 main groups identified from Soria's classification [141] (a) and using the 10 important features by ssFCM methodology in (b), by CKM in (c) and by MBIC in (d).

Table 7.15: Survival curve differences using log-rank test [150] based on CKM-10 clustering for all 1076 patients. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

CKM-10 7 subgroups *							3 main groups *		
Cluster	1	2	3	4	5	6	Cluster	1	2
2	0.734						2	*	
3	0.084	0.094					3	*	0.159
4	*	*	0.002						
5	*	*	0.008	0.683					
6	*	*	*	0.140	0.060				
7	*	*	0.001	0.689	0.483	0.276			

* p< 0.001

Table 7.16: Survival curve differences using log-rank test [150] based on MBIC-10 clustering for all 1076 patients. The first line shows p-value based on comparison with all survival curves together using the two different grouping.

MBIC-10 7 subgroups *							3 main groups *		
Cluster	1	2	3	4	5	6	Cluster	1	2
2	0.591						2	*	
3	0.052	0.090					3	*	0.334
4	*	*	0.001						
5	*	*	0.001	0.996					
6	*	*	*	0.477	0.461				
7	*	*	*	0.492	0.529	0.883			

* $p < 0.001$

- End of thesis -