Developing and Improving Methods for Robust Ensemble Classification: An Aggregation Operator and Clustering-Classification Approach



Utkarsh Agrawal

4246648

School of Computer Science

University of Nottingham

This thesis is submitted for the degree of

Doctor of Philosophy

September 2019

I would like to dedicate this thesis to my family.

Declaration

The work in this thesis is based on research carried out at the Intelligent Modelling & Analysis Research Group and the Lab for Uncertainty in Data and Decision Making (LUCID), in the School of Computer Science at the University of Nottingham, United Kingdom.

Utkarsh Agrawal
4246648
September 2019

Acknowledgements

This thesis is dedicated to my mom, dad, brother and whole family in India and USA, for their unconditional love, unfailing support and encouragement for striving my aspirations throughout my life.

Firstly, I would like express my sincere gratitude to supervisor Dr Christian Wagner for providing me continuous support, guidance and invaluable feedback throughout my studies. I am also thankful to him for plenty of opportunities he gave me for my development, not only academically but personally as well.

I am also grateful to Dr Daniele Soria for giving me the opportunity for my Ph.D. in Nottingham and his constant support throughout my studies. My sincere gratitude also goes to Prof. Jon Garibaldi for his guidance and encouragement. A special mention to Dr Grazziela Figueredo who has always been ready to help me. I am extremely thankful to her for never ending support and always being available for a chat.

Many thanks to the amazing members of IMA and LUCID groups, specifically Direnc, Dr Javier, Shaily, Dr Zack, Dr Josie, Divish and Dr Tuong, for all of the insightful discussions, moments of laughter and vent, and their support in making my studies at the University of Nottingham so productive and fulfilling. Many thanks to Pritesh, Maliha, Srabani, Meezan Bansari, Vivek, Akash, Amit and Tamal who made my life enjoyable in Nottingham. A special mention to my dear friends Namrata and Nitin for always being there. If I forgot to mention someone, you know who you are!

Finally, I would like to thank University of Nottingham and School of Computer Science for financially supporting my studies.

Abstract

Classification is an important technique (in pattern recognition) to categorise objects within their respective groups. In most real world pattern recognition problems, it has become difficult to achieve best performance using an individual classifier. Ensemble algorithms, which are methods combining multiple individual classifiers, have already earned widespread approval within the machine learning community due to their ability to produce results in a wide range of applications. However, some challenges still exist in order to achieve a robust classification, in particular, classification of data points which are difficult to be assigned in one of the groups, and leveraging of existing external knowledge in order to better combine individual classifier outputs (fusion step of the ensemble). This thesis comprehensively explores these two key aspects and issues among the ensemble methods.

The first challenge to generate a robust ensemble classification method is to classify data points which are difficult to label, across the applications using unlabelled datasets (and ensemble clustering frameworks). One specific problem due to this unclassified data is incomplete representation of the dataset. This limitation presents the need to introduce a new framework, which might help to improve the final classification by assigning more data to one of the groups. In this thesis, a robust *two step* framework is presented, which incorporates an *ensemble classification* stage after an ensemble clustering stage. Together, these combine to effectively identify core groups, distribute data within these groups and improve final classification through re-classifying unclustered data (that would otherwise be unassigned to any of the groups). Practical impact of the presented framework is demonstrated through application to novel real world datasets including two breast cancer datasets (breast cancer biological group stratification from the Nottingham and Edinburgh datasets), one heavy goods vehicle dataset (driving stereotypes from the Microlise dataset) and multiple standard datasets from the UCI repository (to demonstrate the robustness of the framework). Results obtained from these datasets show that our novel framework offers an improved, reliable and robust classification technique. These findings were also verified with statistical tests,

visualisation techniques, cluster quality assessment and interpretation from experts (ground truth in case of the UCI repository).

The second challenge focused in this thesis is leveraging external information for improving fusion step of the ensemble for a better ensemble classification performance. Insight on data offers the potentially extremely valuable prospect of leveraging external information. The use of this additional knowledge can lead to better '*ensemble*' classification methods. One approach to capture this information is the use of aggregation operators, which combine the information from multiple sources with respect to a Fuzzy Measure (FM), which captures the worth of all the individual sources and all of their possible combinations. Several approaches to design the FMs exist in the literature; however, these methods do not leverage the external information, which could allow us to better understand the method of data fusion (or ensemble, in the case of ensemble classification). In this thesis, the concept of so called '*A Priori*' FMs is introduced, which are generated based on external information and thus provide an alternative to the existing FM approaches (such as the *algorithm-driven* FMs). The thesis then proceed to develop two specific instances of such an *A Priori* FM in order to support the decision level fusion step in the ensemble classification. This new ensemble classification method is empirically assessed through application to multiple independent datasets. Results indicated that in *cases where external information was available*, the proposed '*A Priori*' FM based ensemble classifier is a robust method achieving improved performances.

Table of contents

List of figures **xvii**

List of tables **xxi**

Nomenclature **xxv**

Dissemination **xxvii**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Aim and Objectives of the Thesis 4

1.3 Contributions to Knowledge 4

1.4 Organisation of the Thesis 5

2 Background **7**

2.1 Classification Algorithms 7

2.1.1 Artificial Neural Networks 8

2.1.2 Nearest Neighbour 10

2.1.3 Support Vector Machines 11

2.2 Ensemble Classification 12

2.2.1 Why ensemble algorithms? 13

2.2.2 Factors affecting Ensemble Algorithm Construction 14

2.2.3 State-of-the-art ensemble classifiers 19

2.3 Clustering Algorithms 21

2.3.1 Hierarchical Clustering Algorithm (HCA) 21

2.3.2 K-Means (with HCA) 23

2.3.3 Partitioning Around Medoids 23

2.3.4 Ensemble Clustering 24

2.3.5 Cluster Validity Indices 24
2.4 Aggregation Functions 26
2.4.1 Fuzzy Measures 26
2.4.2 Aggregation Operators 29
2.4.3 Aggregation operator based ensemble classifiers 31
2.5 Ensemble Classification Applications 34
2.5.1 Breast Cancer Biological Groups Stratification 34
2.5.2 Heavy Goods Vehicle Driving Stereotypes 41
2.5.3 Aggregation Operators in Face Recognition Applications 44
2.5.4 UCI machine learning repositories 45
2.6 Summary 45

3 Combining Clustering and Classification to Enhance Classification of Unlabelled Data 47
3.1 Background and Motivation 48
3.2 A case study illustrating challenges in current classification frameworks . . 49
3.2.1 FuzzyQSBA Algorithm 51
3.2.2 Dataset Description 51
3.2.3 Experiment 52
3.2.4 Results 54
3.2.5 Discussion 56
3.2.6 Summary 57
3.3 Two-stage Data Classification Framework 58
3.3.1 Data Gathering 59
3.3.2 Data Pre-processing 59
3.3.3 Coupling ensembles of Clustering and Classification 59
3.3.4 Further Analysis of the Groups Formed 63
3.4 Summary 64

4 Validating Robustness of the Classification Framework 65
4.1 Case Study 1: Biological Group Stratification in Edinburgh Breast Cancer Dataset 65
4.1.1 Problem Overview 65
4.1.2 Dataset Description 66
4.1.3 Results 67

4.1.4 Clinical Assessment . . . 71
4.1.5 Discussion . . . 73
4.1.6 Summary . . . 73
4.2 Case Study 2: Characteristic Group Stratification in Benchmark Datasets . . 74
4.2.1 Problem Overview . . . 74
4.2.2 Dataset Description . . . 75
4.2.3 Results and Discussion . . . 75
4.2.4 Summary . . . 76
4.3 Case Study 3: Identifying Heavy Goods Vehicle Driving Styles in the United Kingdom . . . 77
4.3.1 Problem Overview . . . 77
4.3.2 Background and Motivation . . . 77
4.3.3 Dataset Description . . . 78
4.3.4 Classification framework tailored to Heavy Goods Vehicle dataset . 80
4.3.5 Driver Profile Labelling Algorithm . . . 85
4.3.6 Experimental Settings . . . 85
4.3.7 Results . . . 86
4.3.8 Discussion . . . 99
4.3.9 Summary . . . 104
4.4 Overall Summary . . . 104

5 Combination of Classifiers Through Leveraging External Knowledge 107
5.1 Background and Motivation . . . 108
5.2 *A Priori* Fuzzy Measure . . . 109
5.2.1 Algorithm to convert external information to Fuzzy Measure . . . 110
5.2.2 Classifier Performance Driven *A Priori* Fuzzy Measure Learning . . 112
5.2.3 Literature Driven *A Priori* Fuzzy Measure Learning . . . 112
5.2.4 *A Priori* Fuzzy Measure Driven Ensemble Classifier . . . 113
5.3 Aggregation operator selection . . . 113
5.3.1 Decision-level Fusion of multiple kernels using Recursive Average (DFRAV) . . . 115
5.3.2 Results and Discussion . . . 117
5.4 Summary . . . 118

6 *A Priori* FM driven ensemble classifiers **121**
6.1 Case Study 1: Generating *A Priori* Fuzzy Measures from Knowledge for Face Recognition . . . 121
6.1.1 Problem Overview . . . 121
6.1.2 Experiment . . . 122
6.1.3 Classifier Performance Driven *A Priori* Fuzzy Measure Learning: CNN experiment based accuracy . . . 123
6.1.4 Literature Driven *A Priori* Fuzzy Measure Learning: Psychological experiments based information values . . . 127
6.1.5 Discussions . . . 129
6.1.6 Summary . . . 129
6.2 Case Study 2: *A Priori* Fuzzy Measures for multiple UCI datasets . . . 130
6.2.1 Problem Overview . . . 130
6.2.2 Datasets and Pre-processing . . . 130
6.2.3 Results and Discussion . . . 131
6.2.4 Summary . . . 132
6.3 Case Study 3: Comparison of Aggregation Operator-Fuzzy Measure based Ensemble Algorithms with State-of-the-art Ensemble Algorithms . . . 135
6.3.1 Problem Overview . . . 135
6.3.2 Background and Motivation . . . 136
6.3.3 Experiments . . . 137
6.3.4 Results and Discussion . . . 137
6.3.5 Summary . . . 139
6.4 Overall Summary . . . 140

7 Conclusions **143**
7.1 Contributions . . . 143
7.2 Thesis Summary . . . 144
7.3 Limitations . . . 147
7.4 Future Works . . . 148

References **151**

Appendix A Ki67 importance in Breast Cancer Stratification **165**
A.1 Introduction . . . 165
A.2 Background . . . 166

A.2.1 Reverse Phase Protein Microarray 166
A.2.2 Patient Samples 167
A.3 Data pre-processing 167
A.4 Methodology 168
A.5 Results 168
A.5.1 Analysis on 10 biomarkers 168
A.5.2 Analysis on 11 biomarkers 170
A.5.3 Analysis on 14 biomarkers 171
A.5.4 Discussion 172

List of figures

2.1 Single Perceptron representation (adapted from [1]) 8
2.2 A representation of ANN (replicated from [2]) 10
2.3 Nearest Neighbour classification representation (replicated from [3]) 11
2.4 SVM decision planes (replicated from [4]) 12
2.5 Pictorial representation of the reasons why ensemble methods may work better than a single classifier (replicated from [5]) 14
2.6 Ensemble connection Topology 15
2.7 Kappa-error diagram for five ensemble methods on 31 datasets with 11 base classifiers for model represented in left image and 1000 base classifiers for model represented in the right image (replicated from [6]) 16
2.8 Parallel connection model 17
2.9 Serial connection model 18
2.10 Hierarchical Clustering Algorithm visual representation 22
2.11 FM Lattice for three sources 27
2.12 RAV for three sources (replicated from [7]) 31
2.13 Breast Cancer Classification structure by Soria et al. [8]. 37
2.14 Subdivision of HER2+ in two subclasses HER2+/ER+ and HER2+/ER- based on immunohistochemical profiles by Green et al. [9]. 38
2.15 Breast Cancer Classification structure by Soria et al. [10]. 40

3.1 Flowchart of the research on Nottingham Breast Cancer Dataset 50
3.2 Tuples of data set produced after applying the FuzzyQSBA algorithm. . . . 52
3.3 Boxplot of 1073 patients from the Nottingham Breast Cancer Series. 57
3.4 Boxplots of patients (a) Class4 of original classification (b) Class4 of new methodology (c) Class5 of original classification (d) Class5 of new methodology. 58
3.5 Abstract steps in the proposed framework 59

3.6 Flowchart of the sub-steps in 'Coupling Clustering and Classification to form groups' step from the abstract view. 'Ensemble clustering (Clusterfusion) to form groups' and 'Ensemble Classification approach to classify more data' are further subdivided into smaller steps explained in Figures 3 and 4 respectively. 60
3.7 Flowchart for Consensus Clustering Unit shows the pictorial representation of the order of steps to form groups in a dataset and subsequently divide the data in these constituted groups. 62
3.8 Flowchart for the order of steps in Consensus Classification Unit to classify more data in groups constituted in the clusterfusion step. 63

4.1 Boxplots of 478 patients in the Edinburgh Dataset 66
4.2 Boxplots of biomarkers. (a), (c), (e), (g): for patients in clusters derived after ensemble clustering. (b), (d), (f), (h): for additional patients classified after ensemble classification. 69
4.3 Breast Cancer Groups with representative characteristics 72
4.4 Flowchart of the framework to determine the clusters in the drivers data set 81
4.5 Data set features density distribution 82
4.6 Example of the comparison of each variable with the median values found in the cluster. The corresponding labels are on the right side of the figure. . . . 85
4.7 Validity indices for (A) K-means and (B) PAM for the instances with low daily average mileage 87
4.8 Clusters within low daily average distance 89
4.9 Clusters within low daily average distance: comparing each variable with the median values found in the cluster. 90
4.10 Validity indices for (A) K-means and (B) PAM for the instances with medium daily average mileage 92
4.11 Clusters within medium daily average distance 94
4.12 Clusters within medium daily average distance: comparing each variable with the median values found in the cluster. 95
4.13 Validity indices for (A) K-means and (B) PAM for the instances with high daily average mileage 96
4.14 Clusters within high daily average distance 98
4.15 Clusters within high daily average distance: comparing each variable with the median values found in the cluster. 99

4.16 Scatter plot of the drivers in their respective driving profiles for (a) low, (b) medium (c) high daily average distance and (d) the entire data set 102

5.1 Proposed Fuzzy Measure Construction Framework. 110
5.2 FM values comparison of CFI and RAV 119

6.1 Sample image from AT&T dataset (replicated from [11]) 123
6.2 Fuzzy Measure Lattice for classifier performance driven accuracy 125
6.3 Fuzzy Measure Lattice for literature knowledge driven accuracy 128

A.1 Boxplot of the RPPA data 167

List of tables

2.1 Rules associated with validity indices 24
2.2 Driving behaviour clusters from Constantinescu *et al.* [12] 43

3.1 Results after classification 55
3.2 Results after clustering 56

4.1 Number of clusters suggested by different validity indices 67
4.2 Division of patients in clusters 68
4.3 The p-value comparison for patients after ensemble clustering with ensemble classification 70
4.4 DB index value of each subgroup of patients assigned after ensemble classification, compared by assigning to all the other remaining clusters. 71
4.5 Descriptive features such as number of classes, number of features and total data in the benchmark datasets 75
4.6 A comparison of distribution of data among the groups for the benchmark datasets. 76
4.7 Three main subgroups based on average distance travelled 86
4.8 Distribution of drivers among the groups for Low Daily Average Distance . 88
4.9 p-value of majority voting for Low Daily Average Distance 88
4.10 Profiles for Low Daily Average Distance 91
4.11 Distribution of drivers among the groups for Medium Daily Average Distance 91
4.12 p-value comparison of ensemble clustering with ensemble classification for Medium Daily Average Distance 93
4.13 Profiles for Medium Daily Average Distance 93
4.14 Distribution of drivers among the groups for High Daily Average Distance . 97
4.15 p-value comparison of ensemble clustering with ensemble classification for High Daily Average Distance 97

4.16 Profiles for High Daily Average Distance 97
4.17 Driving Profiles 101

5.1 Comparison statistics of DeFIMKL and DFRAV. For clarity, the best classifiers are bold and the classifiers performances which are statistically significantly indifferent than the best are in *italics* 118

6.1 Observed *A Priori* performance from the two experiments 124
6.2 Accuracy (in %age) comparison of the classifier performance driven A Priori FM with Uriz and Sugeno λ-measure for both the CFI and the RAV 126
6.3 Accuracy (in %age) comparison of the literature driven A Priori FM with Uriz and Sugeno λ-measure for both the CFI and the RAV 129
6.4 Dataset Description 131
6.5 Summary performance statistics of the *A Priori* FM, the Uriz FM and the Sugeno λ-measure based ensemble classifiers for the same base classifiers. For clarity, the best classifiers are bold, the classifiers performances which are statistically significantly indifferent than the best are in *italics* and the ones which are statistically significantly worse than the best are underlined. 133
6.6 Summary performance statistics of the *A Priori* FM, the Uriz FM and the Sugeno λ-measure based ensemble classifiers with mixed classifiers. For clarity, the best classifiers are bold, the classifiers performances which are statistically significantly indifferent than the best are in *italics* and the ones which are statistically significantly worse than the best are underlined. . . . 134
6.7 Summary performance statistics of the *A Priori* FM based ensemble classifier comparison with state-of-the-art ensemble classifiers. For clarity, the best classifiers are bold, the classifiers performances which are statistically significantly indifferent than the best are in *italics* and the ones which are statistically significantly worse than the best are underlined. 138

A.1 Results after division in three classes using 10 biomarkers 169
A.2 Results after division in seven (and NC) classes using 10 biomarkers 169
A.3 Results after combining with classifusion using 10 biomarkers 170
A.4 Results after division in three classes using 11 biomarkers 170
A.5 Results after division in seven (and NC) classes using 11 biomarkers 171
A.6 Results after combining with classifusion using 11 biomarkers 171
A.7 Results after division in three classes using 14 biomarkers 171

A.8 Results after division in seven (and NC) classes using 14 biomarkers 172
A.9 Results after combining with classifusion using 14 biomarkers 172

Nomenclature

ANN Artificial Neural Network

CFI Choquet Fuzzy Integral

CI Computational Intelligence

CNN Convolutional Neural Networks

DB Davies-Bound

DeFIMKL Decision-level Fuzzy Integral Multiple Kernel Learning

FI Fuzzy Integral

FM Fuzzy Measure

HCA Hierarchical Clustering Algorithm

HGV Heavy Goods Vehicle

MJSVM Majority Voting with SVM

MKL Multiple Kernel Learning

ML Machine Learning

MLP Multi-Layer Perceptron Neural Network

MRE-CNN Multi-Region Ensemble Convolutional Neural Network

NN k-Nearest Neighbour

NPI Nottingham Prognostic Index

PAM Partitioning Around Medoids

PCA Principal Component Analysis

QP Quadratic Programming

QSBA Quantifier Subsethood-based Algorithm

RAV Recursive Average

SVM Support Vector Machines

WSBA Weighted Subsethood-based Algorithm

Publications

P1 Utkarsh Agrawal, Daniele Soria, Christian Wagner. '*Cancer Subtype Identification Pipeline: A Classifusion Approach*." in IEEE Congress on Evolutionary Computation (CEC) at World Congress on Computational Intelligence (WCCI), pp. 2858-2865, 2016.

P2 Utkarsh Agrawal, Daniele Soria, Christian Wagner, Jonathan Garibaldi, Ian O. Ellis, John M. S. Bartlett, David Cameron, Emad A. Rakha, Andrew R. Green. '*Combining Clustering and Classification Ensembles A Novel Pipeline to Identify Breast Cancer Profiles*.' in Elsevier's Artificial Intelligence in Medicine Journal (AIIM), pp. 27-37, Vol. 97, doi 10.1016/j.artmed.2019.05.002, 2019.

P3 Grazziela P. Figueredo, Utkarsh Agrawal, Jimiama M. Mase, Mohammad Mesgarpour, Christian Wagner, Daniele Soria, Jonathan M. Garibaldi, Peer-Olaf Siebers, Robert I. John. '*Identifying Heavy Goods Vehicle Driving Styles in the United Kingdom*.' in IEEE Transactions on Intelligent Transportation Systems (ITS), pp. 3324-3336, Vol. 20, Issue 9, doi 10.1109/TITS.2018.2875343, 2018.

P4 Utkarsh Agrawal, Christian Wagner, Jonathan M. Garibaldi, Daniele Soria. '*Fuzzy Integral Driven Ensemble Classification using A Priori Fuzzy Measures*.' in IEEE International Conference on Fuzzy Systems (FUZZ), 2019.

P5 Utkarsh Agrawal, Anthony J. Pinar, Christian Wagner, Timothy C. Havens, Daniele Soria, Jonathan M. Garibaldi. '*Comparison of Fuzzy Integral-Fuzzy Measure based Ensemble Algorithms with the State-of-the-art Ensemble Algorithms*.' in International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), pp. 329-341, 2018.

Chapter 1

Introduction

In order to construct a robust and efficient ensemble classification method, this thesis focuses on two key challenges within classification: *to tackle difficult to classify data points* across unlabelled datasets and *to leverage existing external information for improving the fusion step* of ensemble classification methods. Section 1.1 describes the background and motivation for the research in this thesis. Section 1.2 then presents the aims and objectives of this thesis, followed by the contribution to knowledge in Section 1.3. Lastly, Section 1.4 presents the thesis structure.

1.1 Background and Motivation

Classification is a pattern recognition technique employed to determine the group in which a new data point belongs [13]. A classifier uses a training set to learn the classification rules to assign class labels to these data points. Most real-world pattern recognition problems have become very difficult to solve with a single (individual) classification model. Moreover, several classification algorithms exist in the literature having different methods and producing dissimilar results. Ensemble classification methods, which are methods combining multiple individual classification algorithms for an overall superior classification performance [14], have proven to be efficient in these scenarios.

Recently, ensemble classification methods have become popular in the machine learning community and have found application in a wide range of problems including remote sensing [15], finance [16], computer vision [17], pattern recognition [18] and bioinformatics [19]. A study by Bashir *et al.* [20] suggest that ensemble methods mimic human reasoning as they consider multiple opinions before making a final decision, suggesting they could improve the

overall classification performance, one reason for their popularity among machine learning researchers. These claims were analysed in an independent study by Brown *et al.* [14], comparing error distribution of unbiased individual classifiers & unbiased ensemble classifier with biased individual classifiers & biased ensemble classifiers. The work concluded that, overall ensemble classifiers had better classification performance than individual ones, in both the biased and the unbiased cases.

Although ensemble classification methods have been extensively researched, some drawbacks still exist in order to establish a robust classification method. Specifically, two key issues among the ensemble methods are focused in this thesis; classification of data points which are difficult to classify and improvement in understanding the configurations of the fusion step in the ensemble by leveraging external available information.

The first key challenge within the ensemble classification method is classification of instances which are difficult to classify, specifically within applications employing unlabelled datasets. Assigning class labels is important to effectively and efficiently abstract important underlying informations within these datasets. For example, stratification of patients within the biological groups across Breast Cancer datasets and driving stereotypes within Heavy Goods Vehicle (HGV) datasets. To address the classification challenges within the unlabelled datasets, clustering algorithms based classification frameworks have attracted a lot of attention among researchers [8]. Similar to ensemble classification, it is advantageous to use several clustering techniques and combine there results to form one superior algorithm, called 'ensemble clustering'. Ensemble clustering algorithms have also found numerous applications including medicine [8], image segmentation [21] and driving pattern identification [22].

Although ensemble clustering algorithms can elucidate characteristic groups from these datasets, they cannot assign labels to data which are difficult to classify. Previously, Soria et al. [8, 10] used ensemble clustering to construct a classification framework which was able to identify core biological classes within a Breast Cancer dataset and subsequently distribute patients across them. The work divided a number of patients (data) among seven biological groups, however the methodology was unable to classify patients which were difficult to classify, leaving them as 'unclassified'. In another independent analysis, employing ensemble clustering framework identified six driving stereotypes amongst the UK HGV driving community by analysing a telematics dataset of driving incidents for 21,193 drivers [22]. 18,784 drivers were assigned to one of the six classes, leaving behind 2,409 which were difficult to classify. Therefore, in this thesis a new '*two staged*' framework is introduced, which couples the ensemble classification stage after the ensemble clustering

stage; together resulting in an improved classification as they are able to classify these difficult data points (explained later in chapters 3 and 4).

The second key challenge (to achieve better classification performance) is to construct an improved decision level fusion ensemble classifier. One approach to achieve this is to leverage an additional level of insight from the data. Recently, aggregation operators have been used to capture this insight from data, which are mathematical functions that enable the fusion of information (evidence) from multiple sources in respect to a Fuzzy Measure (FM) - which captures the worth of both the individual sources and all their possible combinations [7, 23]. In the context of ensemble classification, aggregation operators together with FMs extend the concept of weighted average ensembles and are able to capture the interactions among the classifiers in the ensemble, resulting in a non-linear ensemble classifier.

Defining FMs in a real world task is challenging, even for a relatively small number of sources. Several approaches to designing FMs exist in the literature, however they suffer from some drawbacks or cannot leverage the external information. FMs learned from training data are constrained by both the selected aggregation operator and the information content of training set. Strategies to derive algorithmic FMs do not capture the true interactions and are commonly driven by mathematically required properties, rather than application-specific insight. Lastly, it is difficult to manually specify FMs, for example using expert knowledge, as the number of required parameters increases exponentially with the number of sources (more details on these drawbacks in Chapters 5 & 6).

This thesis explores ways of addressing the aforementioned drawbacks. In particular, this thesis parameterises the *a priori* knowledge from external information to distil the appropriate FM for a given combination of sources, and subsequently puts forward a fourth category of FM, the so-called '*a priori* FMs'. Two different pathways to capture this *a priori* knowledge are examined: 1) Classifier (Machine Learning) driven, and 2) Literature knowledge driven. Using the external information, one specific instance of such an *A Priori* FM is to support the decision level fusion step in ensemble classification. A detailed empirical study of this presented ensemble method can help to understand the details in the fusion step of the ensemble classification. Chapters 5 & 6 in this thesis present more details on the *a priori* FM.

In the next section we present the aim and objectives to tackle the aforementioned challenges (in ensemble classification).

1.2 Aim and Objectives of the Thesis

The aim of this thesis is to improve existing mechanisms underpinning ensemble classification algorithms and to develop a new robust[1] method for efficient ensemble classification. To achieve this aim, the following objectives were identified:

1. To investigate the key issue of difficult to classify data points, specifically across the classification applications using unlabelled datasets.
2. To establish an enhanced and robust classification framework to tackle this challenge (tackling difficult to classify points) across the applications with unlabelled datasets.
3. To establish and develop new ways of combining classifier outcomes for the cases where additional data about the information sources is available.
4. To evaluate the robustness and real-world impact of the proposed classification framework with application to multiple datasets.
5. To evaluate the performance of the proposed external information based ensemble generation method by presenting a comprehensive study which contrasts its results with multiple state-of-the-art methods in the literature.

1.3 Contributions to Knowledge

The research in this thesis improves current and develops new methods for robust ensemble classification. The contributions from this are:

1. A comprehensive study discussing the challenges in the classification of data points which are difficult to classify.
2. Introduce an enhanced and robust framework by adding an '*ensemble classification*' step after the ensemble clustering step which improves the final classification across the unlabelled datasets.
3. Real-world impact of the proposed classification framework with application to multiple datasets, specifically two breast cancer datasets for biological group stratification, one heavy good vehicle dataset for driving stereotype identification and multiple standard datasets from the UCI repository.

[1] herein robust method means an approach that can be modified for different datasets without compromising good results

4. Identifying and discussing key issues with the current Fuzzy Measures (FMs) and subsequently establish a new category of FMs, the '*A Priori* FM'.

5. A comprehensive study contrasting the results of the '*A Priori* FM' (in respect to an FI aggregation operator) based ensemble classification approach with state-of-the-art aggregation operator based ensemble methods (previously proposed in literature) across multiple case studies. In addition, this thesis also contextualises their results with the traditional state-of-the-art ensemble classification algorithms for all the selected case studies.

1.4 Organisation of the Thesis

Chapter 2 presents a literature review of various classification algorithms considered in this thesis and proposed in the past three to four decades. A review of the existing ensemble classification algorithms is then presented, highlighting the strengths and key challenges within these methods. This chapter then investigates the classification using the clustering algorithms. To understand the ensemble generation i.e. fusion step in the ensemble classification, a review of the existing aggregation operators based non-linear ensemble classifiers is also presented. To conclude this chapter, key issues within existing applications are discussed considering the applicaitons of breast cancer, heavy goods vehicle and face recognition.

Chapter 3 first reviews the current literature in respect to unlabelled datasets employing ensemble clustering techniques for classification of data. One case study on Breast Cancer is presented to illustrate the need for an enhanced classification framework. The case study highlights how current method for the categorisation of data in key applications risk leaving a subset of data unclassified. Chapter 3 then builds upon these findings to introduce a novel two-stage classification framework which employs an additional ensemble classification stage after the ensemble clustering stage, for the identification of core groups, data distribution in them and improvement in final classification results by targeting the unclassified data points.

Chapter 4 presents three comprehensive empirical case studies to evaluate and demonstrate the classification framework in real-world applicaitons. These case studies are important as they inform the way we design the theoretical frameworks. Additionally, they also allow to systematically evaluate these frameworks. In the first case study, the framework is employed on Edinburgh breast cancer datasets for biological group stratification, the second study makes use of multiple standard datasets from UCI repository for stratification of their

characteristic groups, and the third case study employs the Microlise's heavy good vehicle dataset for driving stereotype identification.

Chapter 5 presents a novel method for classification for cases where additional external information about the data is available. In order to do so, aggregation operator based fusion is employed. In the context of ensemble classification, aggregation operators together with FMs extend the concept of weighted average ensembles and are able to capture the interactions among the classifiers in the ensemble, resulting in a non-linear ensemble classifier. In this thesis, we define the concept of '*A Priori*' FMs which are generated based on external information and thus provide an alternative to the traditional approaches of FMs. Two different pathways to capture the *a priori* knowledge will be examined: 1) Classifier (Machine Learning) driven, and 2) Literature knowledge driven. In order to select the most appropriate aggregation operator for the *A Priori* FM, the CFI will be compared with the recently introduced RAV operator. This chapter then proceeds to support the decision level fusion step in ensemble classification by employing this *A Priori* FM employing both the CFI and the RAV as possible fusion operators.

Chapter 6 empirically evaluates the characteristic details of the fusion step in ensemble classification, by employing the A priori FM introduced in Chapter 5. First, a detailed case study on face recognition employing the A Priori FM ensemble classifier is presented. Next, a detailed case study comparing the novel A priori FM based ensemble classifier with the Uriz FM and the Sugeno λ-measure based ensemble classifier for both the CFI and the RAV aggregation operator. Lastly, a comprehensive study comparing the A priori FMs based ensemble classifier with state-of-the-art aggregation operator-FM based ensemble classifier DeFIMKL; and four other traditional ensemble classifiers including Adaboost, Bagging, Random Forest and Majority Voting.

Lastly, Chapter 7 summarises and concludes the work presented in this thesis, highlighting the main contributions, articles produced, limitations and future works.

Chapter 2

Background

This chapter will outline the current research gaps and the existing literature on CI methods employed in this thesis. Section 2.1 and 2.2 provides a brief overview of the individual and ensemble classification algorithms employed in this thesis. These sections also discusses the evolution of ensemble methods as the state-of-the-art in pattern recognition. Section 2.3 specifically explores the clustering algorithm frameworks employed for unlabelled datasets. Section 2.4 puts forward Fuzzy Measures and Aggregation operators respectively as tools for performing aggregation. This section also identifies the key gaps in the literature regarding aggregation operator based ensemble classification methods. Lastly, Section 2.5 of this chapter outlines previous research (in particular highlighting gaps) and contributions in the literature for the specific case studies that are used in this thesis.

2.1 Classification Algorithms

Classification, one example of supervised learning, is a technique to determine the group in which a new object belongs by assigning class labels [13]. A classifier is constructed by learning the classification rules from a training set (a previously labelled set of observations). Using these rules, the goal of the classifier is to predict the final class label of the test instances. Classification algorithms have received great attention in the past few decades in a wide range of applications including computer vision [17], pattern recognition [18], bioinformatics [19], natural language processing [24]. The main steps in classification are as follows:

1. Identify the feature set and select the optimal classification algorithm for the selected dataset.

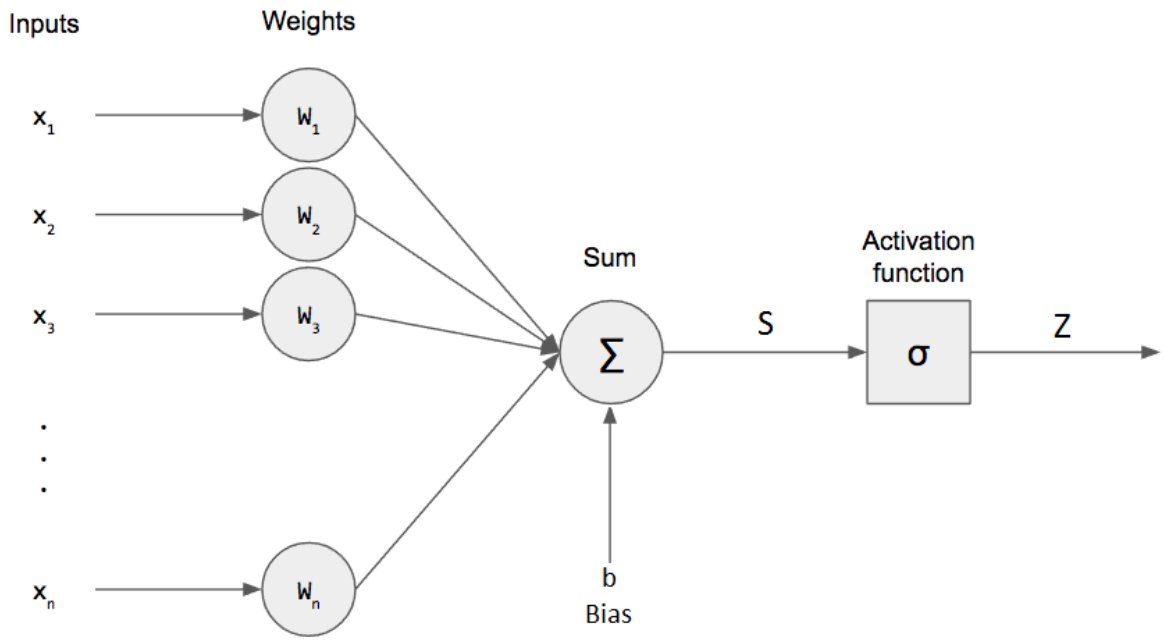


Figure 2.1 Single Perceptron representation (adapted from [1])

2. Divide the dataset into training, validation and test sets.

3. Using techniques such as validation sets and cross-validation, learn the classification rules from the training set.

4. Classify the data from the test set.

5. Evaluate the classification results of the test set.

Although numerous classification algorithms exist in literature, in this thesis, Artificial Neural Networks, k-Nearest Neighbours, Support Vector Machines and Ensemble classification methods are employed, due to their strong performance in the literature. The next sub-sections describe the aforementioned classification algorithms employed in this thesis.

2.1.1 Artificial Neural Networks

Artificial Neural Network (ANN) algorithms are supervised learning algorithms inspired by biological neural networks [25–28]. Recently, numerous ANN models such as Convolutional Neural Network, Deep-Belief Networks, Recurrent Neural Networks, Auto-encoders [29], etc., have been proposed in the literature, with applications in a large number of problems including Facial Recognition, character recognition, speech recognition, forecasting [30, 17,

29], etc. In this thesis, a version of *Shallow Deep Neural Network*, also referred as *Multi-Layer Perceptron (MLP)*, is employed with back-propagation learning method. Figure 2.1 shows a basic model of a single perceptron, where the weighed sum S is computed by multiplying input vectors $x_1, x_2, \ldots, x_n$ with weight vectors $w_1, w_2, \ldots, w_n$ and adding a bias b.

$$S = \sum_{i=1}^{n} (x_i.w_i) + b \tag{2.1}$$

The final value Z is obtained by passing this sum through an activation function σ.

$$Z = \sigma(S) \tag{2.2}$$

Similar steps are repeated for the remaining hidden and output layers, resulting in one final output y. Commonly, the output obtained from above i.e. the actual output (y) is different from the target output (t). This difference is termed as the error (E), given by the following function:

$$E = \sum_{k=1}^{m} (t_k - y_k)^2 \tag{2.3}$$

where, m is the number of output nodes. This error is used to compute the gradient that is needed to calculate the weights of the network. This error gradient is propagated back to update the weights so as to decrease the error, using the backpropogation of error technique introduced by Werbos in 1974 [31]. The next step is to calculate the error gradient given by y', defined below:

$$y'_{jk} = y_{jk}.(1 - y_{jk}) \tag{2.4}$$

$$\delta w_{jk} = -\alpha.y'_{jk} \tag{2.5}$$

where, δw_{jk} is the change in weight between the j^{th} and k^{th} nodes, and α is the rate of learning. After some iterations, when the error has decreased to an acceptable value, training stops.

A fully connected ANN is shown in Figure 2.2. Similar to single perceptron models, the input layer receives a set of inputs from the dataset and transfers the information onto the hidden layer using synapses (Z form the above example). Then, with the help of more synapses the information is passed onto the third layer or the output layer, where a weighted sum is computed and passed to an activation function. Subsequently, the value received after the activation function is compared with a threshold and the error is calculated. Finally, the network learns by back-propagating the error and readjusting the weights, and the learning continues until the error is below the 'error threshold'.

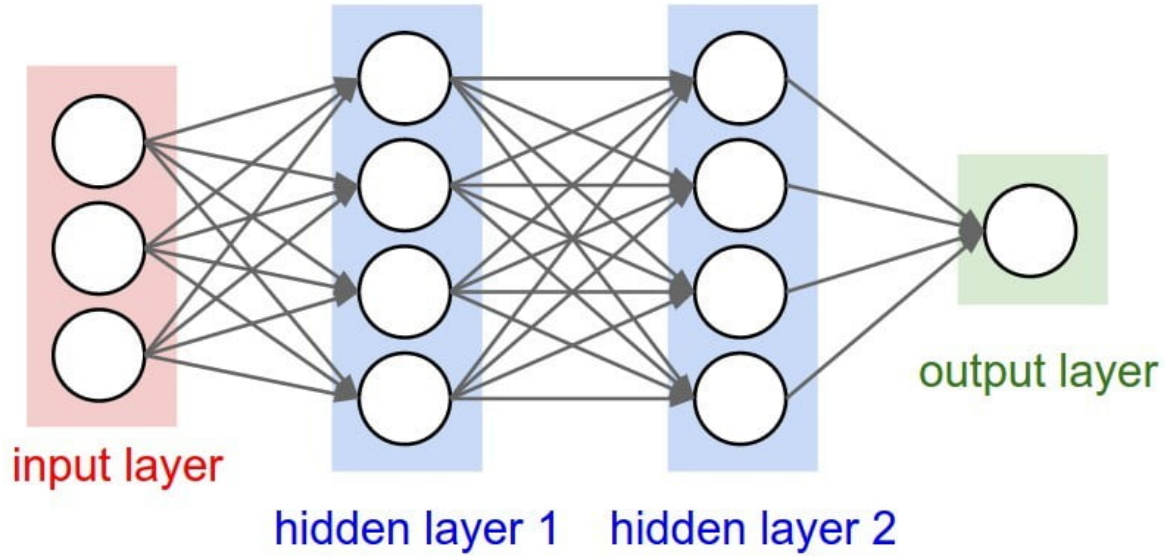


Figure 2.2 A representation of ANN (replicated from [2])

One of the major reasons behind the ANN performing tasks effectively is the structure of its network, its flexibility, its capability to model highly complex models and its non-parametric nature, which enables multiple neurons to solve the problem. Although ANNs have some drawbacks such as determination of network structure, duration of training is unknown, etc. The (aforementioned) advantages massively overshadow its disadvantages, therefore, ANN is currently one of the most commonly used algorithms for classification problems. Due to these aforementioned reasons ANN has been selected as one of the classification models in this thesis.

2.1.2 Nearest Neighbour

The Nearest Neighbour algorithm is a memory based non-parametric model which is employed to classify the data points to the nearest group [32, 33]. The advantages of the Nearest Neighbour over other classification algorithms is its efficient memory handling, as it keeps track of all the classified data [32, 34, 35]. It follows a greedy approach by computing the distance of a test sample with all the data which have been classified in their respective groups. The test samples are then assigned to the nearest group based on the distance matrix calculated in the previous step. The algorithm is represented in Fig 2.3 where x_j is one of the 'test samples'. $\omega 1$, $\omega 2$ and $\omega 3$ represent the identified classes in which the data is intended

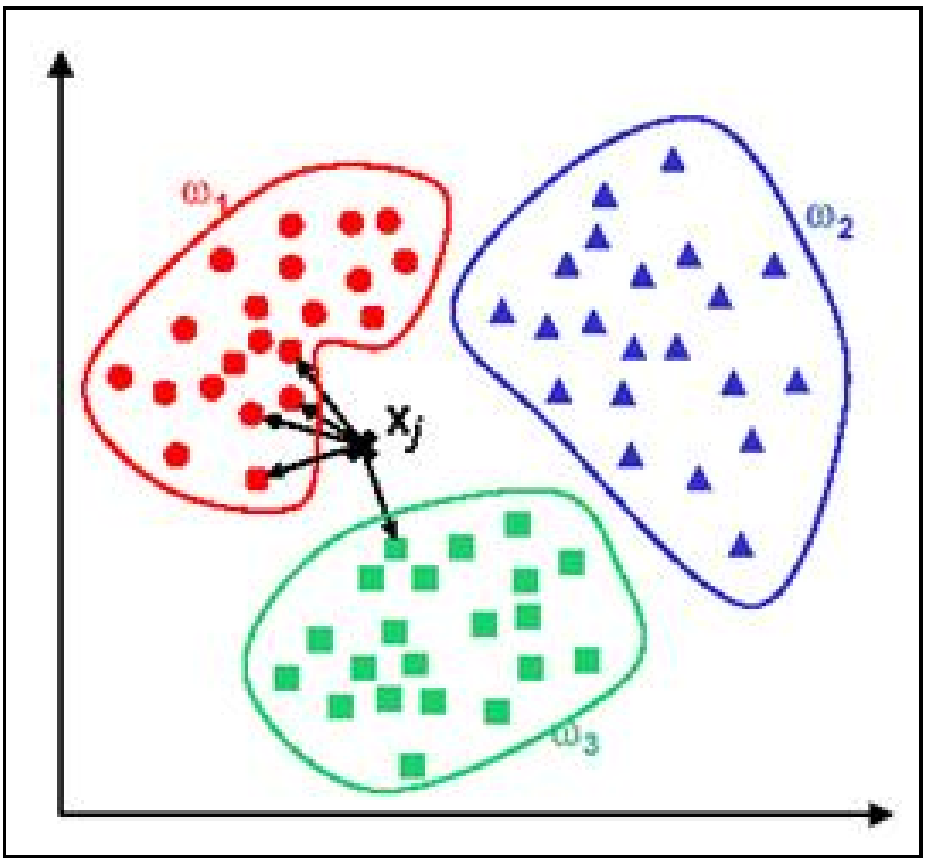


Figure 2.3 Nearest Neighbour classification representation (replicated from [3])

to be classified, computed using an euclidean distance, defined as follows:

$$D(q, p_i) = \sqrt{\sum_{i=1}^{n} (q - p_i)^2} \tag{2.6}$$

$$|P| = min(D) \tag{2.7}$$

where p is the number of mean points of each class and q is the test point to be classified. D is the distance matrix which is used to calculate |P|, which is the point with minimum distance.

2.1.3 Support Vector Machines

Support Vector Machines (SVM)s are classification algorithms based on the concept of decision planes [36]. These planes act as the decision boundary to separate the data. For example, the image on left hand side in Figure 2.4 shows that to divide the data, multiple

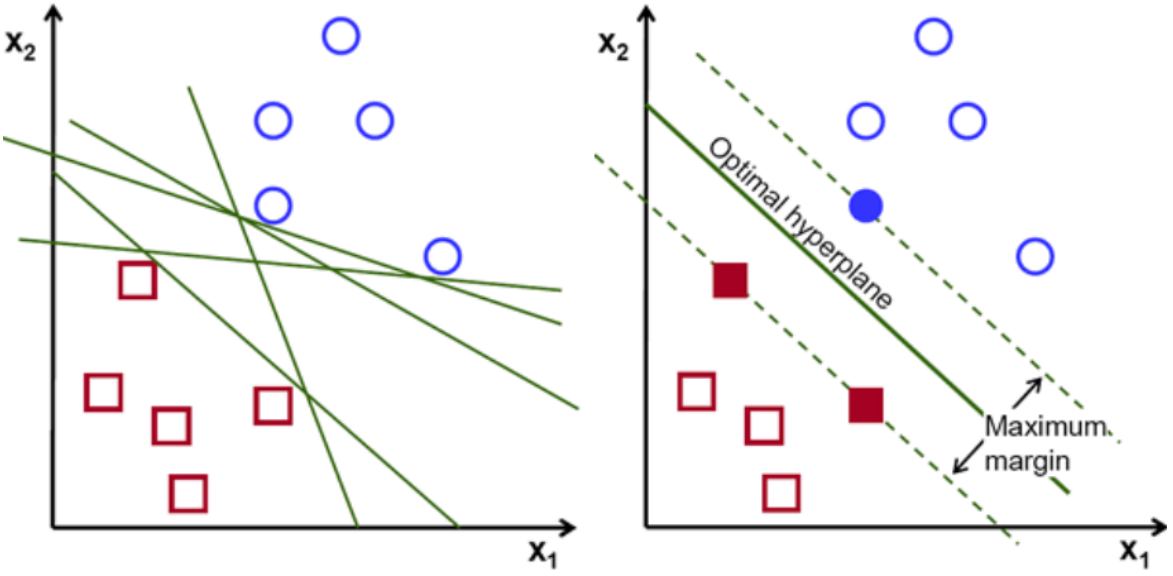


Figure 2.4 SVM decision planes (replicated from [4])

decision planes could exist. SVM works by maximising this distance, by selecting the optimal hyperplane which is equidistant (maximum margin) from the closest data point (support vectors) from both classes, illustrated on the right image in Figure 2.4.

For non-linearly separable data, the concept of kernel-trick is used to separate the data [37]. The kernel-trick transforms the low-dimensional input space to high dimensional space, thus converting the non-linearly separable problem to a linearly separable one. In this thesis, the Radial Basis Function (RBF) kernels is employed, as it is one of the most commonly used kernel in the literature, resulting in good data separation across a large number of problems [38].

The three algorithms i.e. ANNs, Nearest Neighbour and SVMs, are used in multiple ensemble classification models, discussed in the next section.

2.2 Ensemble Classification

Ensemble Classification is a broad term for methods with the objective of classifying data points by combining the outcome of multiple classifiers to improve predictive performance of the system [39, 40]. Previously, several researchers [41–45] have used classifier ensembles to perform classification and discussed their superiority over single classifiers. In the literature, ensemble classification algorithms are also referred to as Meta-classifiers, Multiple Classifiers Systems, Hybrid Intelligent Systems, and Heterogeneous classifiers. In the domain of

supervised learning algorithms, construction of good ensemble classifiers has been an active research area.

An ensemble is constructed by utilising the strengths of individual classifiers, to attain better results by their fusion. Wolpert [46] in 'no free lunch' theorem states: "*There is no single classifier which is optimal for all tasks, since each has capabilities in its own domain*". Before using an ensemble classification algorithm in any problem it is important to understand ***when and why ensemble algorithms are better?***

2.2.1 Why ensemble algorithms?

Krough *et al.* mathematically proved that in order to achieve lowest error, the right balance between diversity and individual accuracy of the classifiers is needed [47]. The proof was further affirmed in a seminal work by Dietterich [5], in which he states that an imperative and sufficient condition for ensembles of classifiers to be more accurate than their independent constituent classifiers are accuracy and diversity. To construct good ensembles, Dietterich [5] further suggests to find the 'hypothesis H' which best generalises the ensemble space. These reasons are explained below with reference to the Figure 2.5. Three fundamental reasons (theory) needs to be addressed (to determine 'H'), explained below:

1. ***Statistical***: The cases when the amount of training data is small, the learning algorithm can find many different hypotheses (h1, h2, h3 h4) that might produce similar accuracy on the training data. For example, in Figure 2.5 (a), the inner curve containing all these aforementioned hypotheses might give good results, although the point 'f' is the true hypothesis. Therefore, using an ensemble the probability to reach optimal approximation might increase.

2. ***Computational***: In some cases it may be computationally difficult to search the best hypothesis. For example, ANN's use the gradient descent strategy to minimise the error and the optimal training time complexity is NP-hard. So, ANNs may get stuck in local optima, far from the best hypothesis. Ensemble classifiers may give better approximation than individual classifiers by combining results of different local searches as shown in Figure 2.5 (b).

3. ***Representational***: In many cases any hypotheses cannot represent the true function. For example, the function 'f' may lie outside, as represented in the Figure 2.5 (c). By combining the hypotheses from more classifiers, the search space can be expanded.

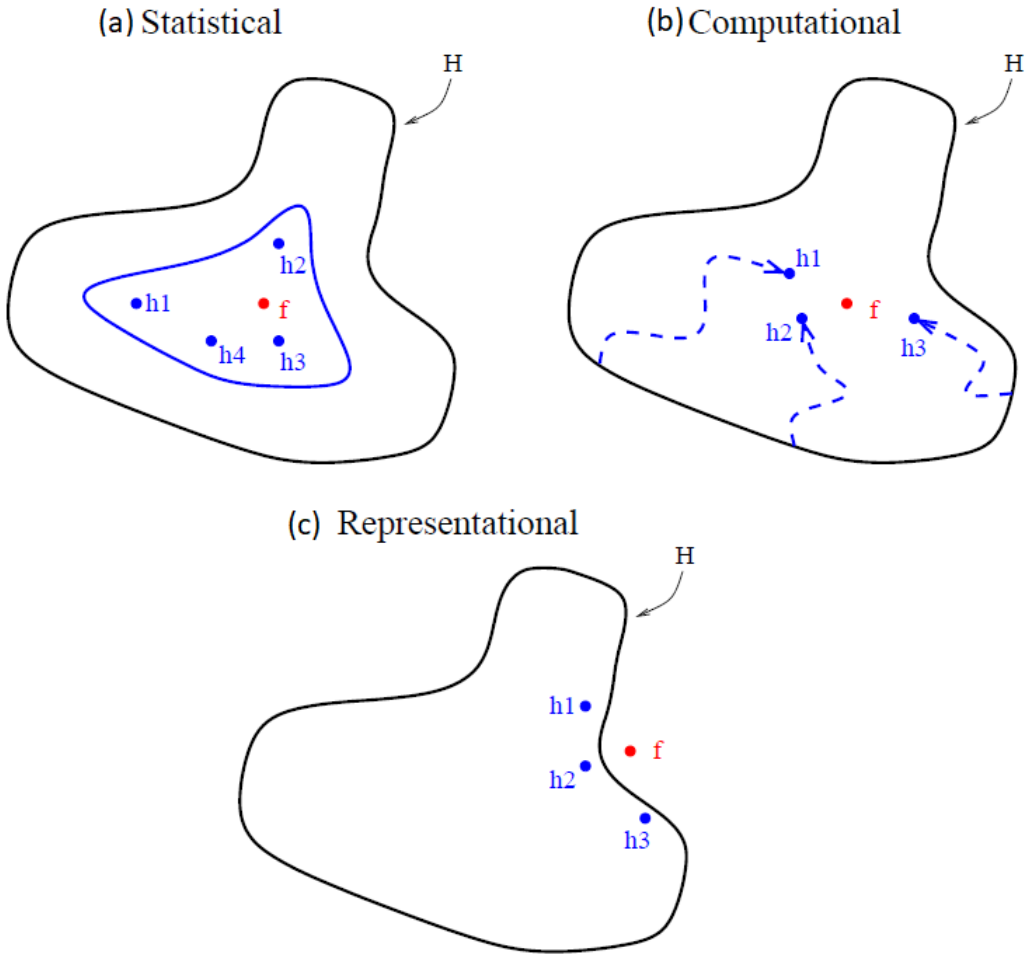


Figure 2.5 Pictorial representation of the reasons why ensemble methods may work better than a single classifier (replicated from [5])

One thing to notice is that the need for this ensemble model occurs in applications with limited training samples.

2.2.2 Factors affecting Ensemble Algorithm Construction

A general architecture of an ensemble classification method is presented in Figure 2.6. An ideal ensemble model combines the classifiers optimally to output superior decisions. To create an appropriate ensemble classifier design, the individual models should cover as much possible search space and at the same time concentrating on minimising the error to provide high accuracy. Two important issues that need to be considered while constructing a superior ensemble are *Classifier Selection* and *Classifier Fusion Design* (described in the next subsections).

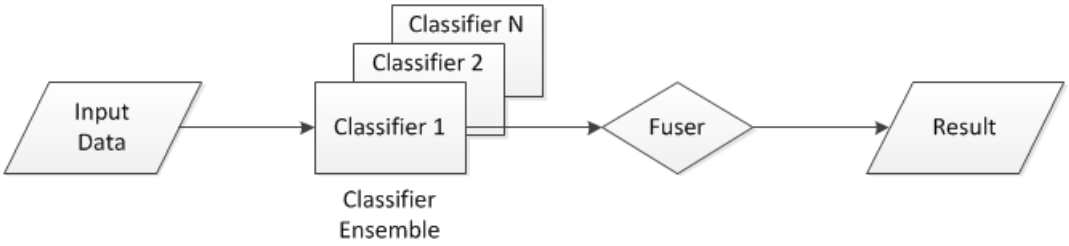


Figure 2.6 Ensemble connection Topology

Classifier Selection

While constructing ensembles, it is important to select classifiers that are mutually complementary, characterised by high diversity and accurate. It is achieved by taking two approaches: implicit or explicit. Implicit methods focus on random techniques based independent classifiers while explicit methods include classifiers from a pool depending upon one another aimed at exploiting strengths of each individual classifier. Some of these factors which ensure diversity within the ensemble are discussed below:

1. ***Diversity Measures***

   As the diversity increases the performance of the ensemble classifier model generally increases [48], although it is important to note that this has not been proven yet. Multiple methods have been proposed in the literature on how to measure the diversity. In an influential work by Sharkey et al., the authors proposed diversity measurement by creating levels of ensemble errors [49]. Four levels of coincident errors among the ensemble members were used to describe the target function. Level 1: Target function is covered and majority vote produces the correct answer (i.e. no coincident error). Level 2: Some coincident error, target function is covered and majority vote produces the correct answer. Level 3: A majority vote will not always yield the right answer. Level 4: The target function is not covered by the ensemble. It is interesting to note that the ensembles at level 2 or 3 can climb to higher level, for example, a level 3 ensemble could contain a subset which might exhibit level 2 and/or level 1 and similarly a level 2 ensemble could contain a subset from level 1.

   In an independent work by Kuncheva [48], they describe pairwise and non-pairwise measure for diversity measurement. Pairwise measures such as Q-static [14], kappa-statistics [50], etc., use a distance metric to determine the diversity using the error coincidence among the classifiers in the ensemble. On the other hand non-pairwise

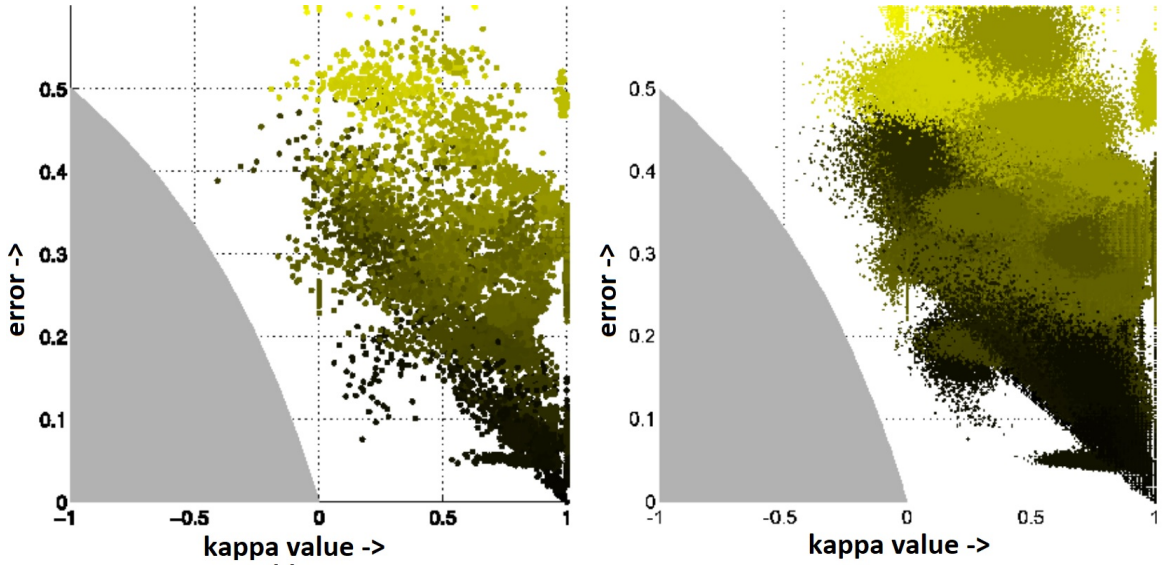


Figure 2.7 Kappa-error diagram for five ensemble methods on 31 datasets with 11 base classifiers for model represented in left image and 1000 base classifiers for model represented in the right image (replicated from [6])

measures focus either entropy or correlation to determine the diversity, for example the Kohave-Wolpert variance [51].

2. ***Diversity of Classifiers***

   One way to ensure diversity is by taking advantage of the bias of the individual classifiers. Lysiak et al. [52] uses competence to create the best ensemble. The probability of correct classification can be used to compare the competence and the probability of misclassification can be used to measure ensemble diversity. Furthermore, the paper proposes Dynamic Ensemble Selection (DES), which constructs ensembles by eliminating weak classifiers and keeping them maximally diverse. Homogeneous ensembles were compared to heterogeneous ensembles, and the results show that diversity has a positive effect on classification quality. Similarly, Pravin et al. [53] discuss the superiority of ensemble methods over individual ones and the need of diversity to create better ensembles. They propose a Classifier Selection Based on Clustering (CSBC), which uses clustering methods to partition base classifiers followed by the selection of classifiers from each cluster to create best ensemble.

   Number of classifiers to be considered. One important factor in diversifying classifiers is the number of models in an ensemble. Kuncheva et al. [6] used Kappa-error diagrams to discuss the trend of diversity of ensembles with accuracy. Figure 2.7 shows the

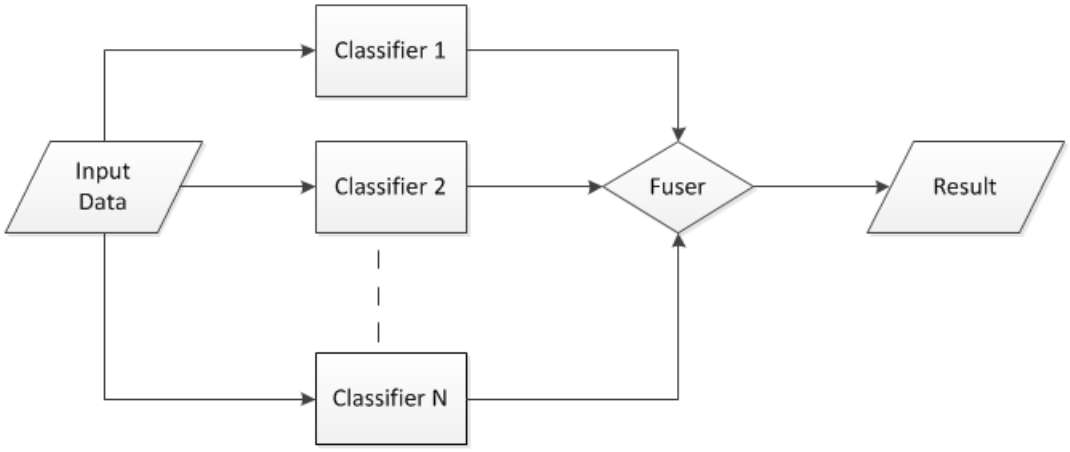


Figure 2.8 Parallel connection model

relation between kappa and error with Y-axis representing the error and X-axis the kappa value. The points are represented for 5 ensemble classifiers on 31 datasets (more details on both in [6]), for 11 and 1000 base classifiers in the left and right model respectively. The figure shows that as the ensemble size increase it becomes more accurate, but at the cost of diversity. For most accurate and diverse ensemble, the goal is to construct an ensemble as closer to shaded area which requires a compromise between error rate and diversity. The algorithm used was: 1) Add the ensemble that increase the performance, 2) Take the average of the predictions and 3) In the end keep the models that increase the performance and reject the others.

3. ***Diversifying Input.*** Different strategies are used to diversify input including data partition and feature set partition.

   Data Partition. A number of reasons lead to data partition include distributed learning in big data problems, data privacy due to external constraints, etc. These models are merged using different ensemble architectures to produce one output.

   Feature Partition. Different classifiers can be trained on different feature spaces and merged at the end for one final output. This strategy could be useful in cases where different classifiers have local optimal solutions [54].

4. ***Diversifying Output.*** Outputs could also be decomposed and obtained in different ways to ensure classification diversity. For example, multi-class classification can be decomposed in to several binary-class problems [54].



Figure 2.9 Serial connection model

Classifier Fusion Design

Another factor in ensemble classification is the classifier topology. The two widely accepted ensemble connection models are parallel and serial. In the parallel topology, each classifier is fed with the same input and the outputs of each classifier are combined in the fuser to obtain the final decision, for example Bagging (discussed in the section 2.2.3). The other used topology is serial which combines classifiers in succession, for example Boosting (discussed in the section 2.2.3). A pictorial representation of the parallel and serial topologies are shown in Figure 2.8 and Figure 2.9 respectively.

Upon deciding upon one of the topology, the classifiers can be fused using one of the following techniques:

1. ***Decision (Class) Level Fusion*** One way to combine the classifiers is to fuse only the outputs (decisions), called as decision level fusion. One of the following three fusion techniques is selected:

   (a) *Unanimous Voting* In unanimous voting mechanism, the test data is assigned an output class if there is consensus among all the considered classification algorithms.

   (b) *Majority Voting* In majority voting mechanism the output is generally chosen either by considering the result with highest votes.

   (c) *Weighted Voting* In weighted voting the output of classifiers are multiplied with weight matrix. Now combine the outputs using either of the majority voting scheme.

2. ***Feature level fusion*** In feature level fusion, different feature sources are combined, usually in multidimensional feature space, to configure one single (or multiple in the case of feature partition diversity measure, see section 2.2.2) input. This method can help classifiers to exploit certain feature in a better way, resulting in (potentially) superior outcomes [23].

3. ***Meta level fusion*** Meta-learning is the process of learning to learn. Informally, a meta-learning algorithm uses experience to change certain aspects of a learning algorithm,

or the learning method itself, such that the modified learner is better than the original learner at learning from additional experience. Some examples of meta learning include Inductive transfer, Self-modifications, and ensemble methods, etc. The most common realisation of meta-learning using ensemble methods are: Stacked Generalisation and Boosting.

2.2.3 State-of-the-art ensemble classifiers

Among a number of ensemble classification methods that exist in the literature [55], five methods are employed in this thesis, described below:

Adaboost

Introduced by Freund et al. [56] in 1997, Adaboost is a very popular ensemble classification algorithm which combines multiple weak classifiers to construct a strong classifier. Adaboost uses a serial topology to combine the outputs of the classifiers. The algorithm trains by giving higher weights to mis-classified data in subsequent iterations (of the classifier) while the weights of the correctly classified instances are decreased, with the objective of correctly classifying these in the next iteration [39]. Hence, after each iteration the weights of the misclassified instances are increased (which was initially equal for all instances) and the weights of the correctly classified instances are decreased. Moreover, depending upon the overall accuracy, an additional weight (higher weight is assigned to more accurate classifiers) is assigned to each individual classifier, which is further used in the test phase. Weighted combinations of all the classifiers in the ensemble are used to predict the final result. The experimental settings and the base algorithm, tailored individually to each problem are discussed in the chapters 4 and 6, where they are used.

Bagging

Bagging (Bootstrap Aggregation) is an ensemble method introduced by Breiman et al. in 1996 [57], which is most commonly used in problems with high variance. Bagging uses a parallel connection topology to combine the outputs of the classifiers. The data for each classifier in the ensemble is selected by sampling with replacement and thus some of the instances may appear more than once in the training set. Each classifier returns the class predictions for the test instances, and the final classification outcome is based on majority-vote. Bagging is effective on nonlinear models such as neural networks and decision trees [58].

Again, the experimental settings and the base algorithm are discussed in the chapters 4 and 6, tailored individually to each problem.

Random Forest

Random Forests are one of the most commonly used ensemble classifiers in the literature. A random forest is a collection of randomised decision trees where each decision tree is learned from different subsets of samples. The random forest classifier, in particular, needs two parameters: the number of classification trees (k), and the number of prediction variables to grow the trees (m) [59]. To classify a test sample each tree is traversed and a vote is assigned to the class based on the probability score. The output is selected by choosing the mode i.e. the output with most votes, of all the k classification outputs. Reducing the number of predictive variables 'm' reduces the correlation between trees, which stops ensemble model from converging to similar generalisation error and in turn helps in increasing the accuracy. Thus, m needs to optimised to minimise the generalisation error.

Majority Voting with SVM (MJSVM)

Let x be an instance and S_i (where $i = 1,2,...k$) be a set of base classifiers (Support Vector Machines, described in the section 2.1.3) that output class labels $m_i(x,c_j)$. For each class label c_j (where $j = 1,...,n$), the output of the final classifier $y(x)$ for instance x is given by:

$$y(x) = \max_{c_j} \sum_{i=1}^{k} m_i(x,c_j). \tag{2.8}$$

More details on the experimental settings of the MJSVM are described in the chapters 4 & 6.

Unanimous Voting

Among a number of decision level fusion methods present in the literature, unanimous voting is chosen for the problem where a high confidence output is necessary, for example in Breast Cancer patients biological group stratification in Chapter 3. The unanimous outcome m among $1,2,..k$ classes for test sample x' can be mathematically represented as:

$$if\{C_1(x'),C_2(x'),...C_n(x')\} = m, \tag{2.9}$$

$$y(x') = m, \tag{2.10}$$

where n is the number of classifiers in the ensemble, $C_n(x')$ is the final decision of the n-th classifier in the ensemble and $y(x')$ if the output class of the ensemble.

The last two sections of this chapter have discussed supervised learning models including individual and ensemble approaches. Next section presents literature on classification using unsupervised learning models.

2.3 Clustering Algorithms

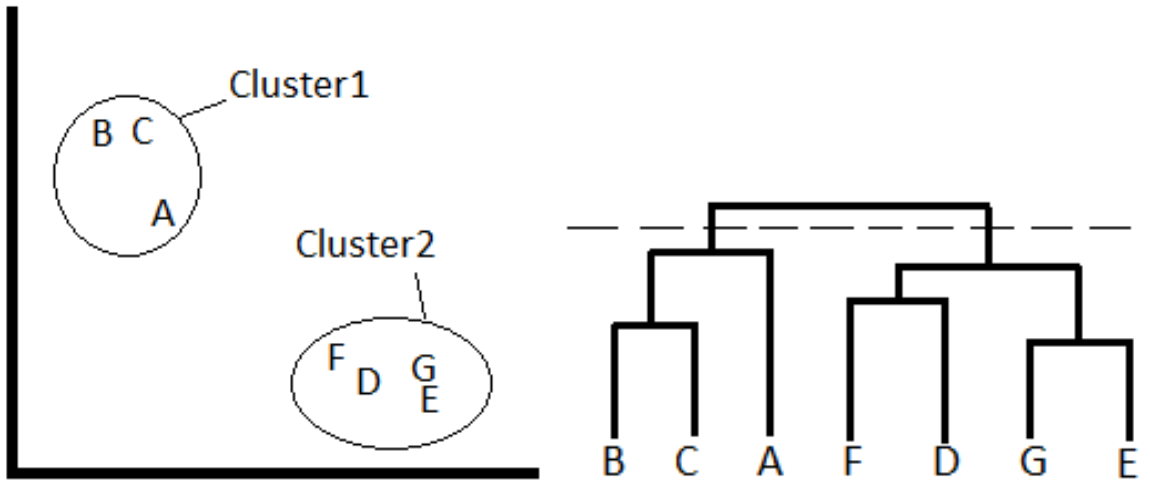
Clustering Techniques or Clustering is an unsupervised Machine Learning (ML) approach, which groups together a set of unlabelled data points (or objects) in a way that the patterns within the same group (cluster) have similar characteristics, and dissimilar to patterns within different clusters. This similarity between points is usually defined using a distance matrix. In the literature, researchers have used these clustering techniques in a vast range of applications including Breast Cancer [8], Heavy Goods Vehicle [60], Recommender Systems [61], etc..

Traditional clustering algorithms are composed of three main steps: 1) Identification of main features, 2) Clustering using one of the aforementioned techniques, and 3) Cluster quality assessment. A wide range of clustering techniques exist such as Hierarchical Clustering, Centroid-based (Partition) Clustering, Density-based Clustering, Grid-based Clustering, Model-based Clustering and Categorical-data Clustering [62–65]. Among many choices, Partitioning Around Medoids (PAM) and Hierarchical Clustering Algorithm (HCA) will be employed as they have traditionally been most commonly used techniques in the applicaitons selected in this work [8, 22, 66]. Traditional clustering algorithms assign each data solely to one cluster, using well defined boundaries. Fuzzy clustering extends the concept of these traditional clustering techniques by assigning each data to every cluster with an associated membership value, leading to undefined boundaries. The two selected algorithms i.e. HCA and PAM define clusters using hard boundaries, contrary to fuzzy clustering methods which define clusters with soft boundaries. Since methods with hard boundaries are selected in this thesis, fuzzy clustering methods are not employed.

The key literature related to the selected clustering techniques are identified and analysed in the following subsections.

2.3.1 Hierarchical Clustering Algorithm (HCA)

Hierarchical Clustering Algorithm (HCA) works by nested grouping of similar data to yield one big cluster [61]. The final output of this algorithm is a ‘dendrogram’, which



(a) Distribution of data within the 2 clusters (b) Dendrogram of these data

Figure 2.10 Hierarchical Clustering Algorithm visual representation

shows a hierarchical relationship between the clusters and the similarity levels at which the groupings change. For example, a dataset with two clusters is shown in Figure 2.10a and its corresponding dendrogram in Figure 2.10b. By vertically positioning the dashed line in the dendrogram, the optimal number (discussed in Section 2.3.5 of this chapter) of clusters can be obtained.

There are two ways to form clusters using HCA, first is agglomerative which follows a bottom up approach; and the second is divisive which follows a top down approach. In this work the agglomerative strategy is used to form clusters as it is less complex then the divisive approach. The algorithm's structure is as follows:

1. Assign each item to a cluster, i.e. if there are n observations then we have n clusters.
2. Find the closest cluster pair based on the pre-specified distance.
3. Merge the two clusters and calculate the average of them, which will be the representation of the new cluster formed.
4. Compute the distance between the new cluster and the old clusters.
5. Repeat the steps 2-4 until all the observations come under one single cluster.

HCA has some limitations including its sensitivity to outliers in the dataset and initial seed selection. However, HCA results in flexible clusters which do need a priori information to cluster data. Moreover, the clusters formed by HCA are only used to find cluster centres to initialise K-Means, defined in the next section.

2.3.2 K-Means (with HCA)

The K-Means clustering algorithm is one of the most commonly used clustering algorithms which works by assigning the data points among k clusters [67, 68]. The steps followed in the algorithm are:

1. Randomly select k number of cluster centres.
2. Compute the Euclidean distance of each observation from all the cluster centres.
3. Assign each observation to the cluster with the minimum distance from the centre.
4. Re-evaluate the cluster centres by considering the means of the respective data points.
5. Repeat steps two to four until no new allocation occurs.

K-Means suffers from the disadvantage of relying on the random selection of cluster centres for each initial run, leading to the formation of dissimilar clusters each time. To tackle this problem multiple solutions have been proposed, and in this thesis we employ the aforementioned agglomerative HCA on the dataset before we run the K-Means algorithm [69]. The tree formed from the HCA is pruned, and the points are passed as the initial cluster centres (now fixed centres) to the K-Means [8]. The K-Means with centres from HCA is the new Modified K-Means (with HCA) algorithm used in this work.

2.3.3 Partitioning Around Medoids

Partitioning Around Medoids (PAM), also known as K-Medoids, is a clustering algorithm which characterises clusters by their medoids (also called the centres) [70]. It is related to both the K-Means and the medoid shift algorithms by minimising the sum of pairwise dissimilarities. At the same time it is dissimilar to the K-Means, as it selects real data points as medoids. The algorithm works in two phases:

1. The build phase, in which a set of k medoids among the n observation points is selected randomly. Then each remaining data point is assigned to one of the closest medoids based on the Euclidean distance.
2. The swap phase, in which each non medoid point is replaced with one of the medoid points and the distance is recalculated. If the new distance is greater in comparison with the previous distance, the swap is reversed, otherwise the process is repeated until all the swaps are performed.

Validity Index	Rule
Marriot	$max_k((d[k+1]-d[k])-(d[k]-d[k-1]))$
Calinski	$min_k((d[k+1]-d[k])-(d[k]-d[k-1]))$
Scott	$max_k(d[k]-d[k-1])$
TraceW	$max_k((d[k+1]-d[k])-(d[k]-d[k-1]))$

Table 2.1 Rules associated with validity indices

2.3.4 Ensemble Clustering

Similar to ensemble classification, ensemble clustering methods merge results of multiple clustering algorithms to form core groups and have been successfully used in the domain of breast cancer [8, 10], image segmentation [64, 65], time series analysis [65], text mining [65], etc. Several researchers have concentrated on contrasting and combining results of different clustering algorithms using re-sampling techniques, combination of outputs, probabilistic methods, pairwise similarity, etc. [71, 72, 65, 64]. Among these multiple ensemble clustering methods, a consensus clustering approach [73] i.e. Clusterfusion is chosen, because it provides robustness, stability and improved solution by increasing the degree of confidence of the groups formed [74].

2.3.5 Cluster Validity Indices

One of the most important factor in cluster analysis is the choice of optimal number of clusters. Cluster validity indices are metrics used to determine the optimal number of clusters by assessing goodness of clustering results [75]. These indices are also helpful in determine the optimal number of clusters in a dataset. In real-world problems, such as the stratification of unlabelled dataset, the number of clusters are not known in advance. By assessing the goodness and quality of clusters, these validity indices are helpful to determine the optimal number of clusters.

The validity indices are either based on compactness of data in a cluster or on the separation of the clusters from one another. A specific rule is associated with each of the indices, as shown in Table 2.1, which specifies the optimal number of clusters k by ranking them. By using these measures in a dataset, the most stable cluster size can be determined. Several validity indices have been proposed in the literature, but in this thesis four commonly used validity indices Marriot, Calinski, Scott and TraceW (d) are used, which are defined in the following subsections [75, 8].

Marriot Validity Index

Marriot Validity Index calculates the cluster quality by minimising the determinant of the intra-cluster co-variance matrix [76]. For k clusters and sum of square W_i within the i_{th} cluster, the marriot index is given by:

$$M = k^2 det(W_i) \tag{2.11}$$

This index is very useful in cases where the data is highly correlated, as no excessive weight to any data.

Calinski Validity Index

Calinski (short for Calinski and Harabasz) Validity Index use inter and intra cluster scatter to compute the goodness of the clusters [77]. For n observations, k number of clusters, z is the centroid of the entire dataset and Z_i being the centroid of the i^{th} cluster, the calinski index is computed as:

$$C = [\frac{\sum_{i=1}^{k} |n_i| ||z_i - z||^2}{k-1}] / [\frac{\sum_{i=1}^{k} \sum_{l=1}^{|n_i|} ||x_l - z_i||^2}{n-k}] \tag{2.12}$$

where $|n_i|$ is number of observations and x_l is the l^{th} observation in the i^{th} cluster.

Unlike other validity indices, Calinski index can be used for all types of data which are represented in the Euclidean space. It has been one of the most commonly used index in the literature [78, 77].

Scott Validity Index

Scott (short for Scott and Symons) Validity Index is based on the assumption that all the data in a sample are taken from a number of different distributions [79]. It is important to note that often little or no information is available about the distributions. Subsequently, the maximum likelihood estimate of the distribution is used to cluster the data. A scatter matrix of log-likelihood of this estimate is used to compute the final Scott index (for more details, refer to [79]). This index is most suitable for model based clustering such as the K-Means algorithm, defined in section 2.3.2.

TraceW Validity Index

Trace of W i.e. TraceW, assess the cluster quality be minimising the within group scatter (intra cluster sum of squared error) [80]. TraceW presents an advantage over other indices

by being invariant under orthogonal transformations, specifically useful for dense clusters. Let, there be a dataset with 'm' number of data points with 'n' dimensional dataset and 'k' number of clusters. For i^{th} group having m' data points represented by P_l, with centre of gravity as C_i, the scatter matrix W_i is given by:

$$W_i = \sum_{l=1}^{m'} (P_l - C_i)^T (P_l - C_i) \tag{2.13}$$

The summation over all the clusters k defines the within group scatter matrix W, given by

$$W = \sum_{i=1}^{k} W_i \tag{2.14}$$

Selecting an appropriate number of clusters

The four validity indices discussed above are either based on dispersion of clusters and between clusters, or consider the scatter matrix of the data points (clusters centres). Each index has distinct properties, therefore an ensemble of these indices is chosen. If the indices suggest different numbers of clusters, then the minimum sum of ranks can be taken as the optimal value. Recently, aggregation operator based ensemble classification have attracted some attention due to their ability to perform non-linear combinations. In the next section, key type of aggregation functions (Fuzzy Measures and aggregation operators) and the classification algorithms based on these aggregation functions are discussed.

2.4 Aggregation Functions

2.4.1 Fuzzy Measures

Fuzzy Measures (FMs) capture the worth of each information source (also called densities) and all their possible combinations i.e. every subset in a power set [7, 23]. Figure 2.11 shows the FM weight structure (also referred as a lattice) for three sources.

Let $X = \{x_1, ..., x_n\}$ be a discrete and finite set of information sources and $g : 2^X \rightarrow [0,1]$ be a FM having the following properties:

P1: Boundary condition, i.e., $g(\emptyset) = 0, g(X) = 1$, and

P2: Monotonic and non-decreasing, i.e., $g(A) \leq g(B) \leq 1$, if $A \subseteq B \subseteq X$.

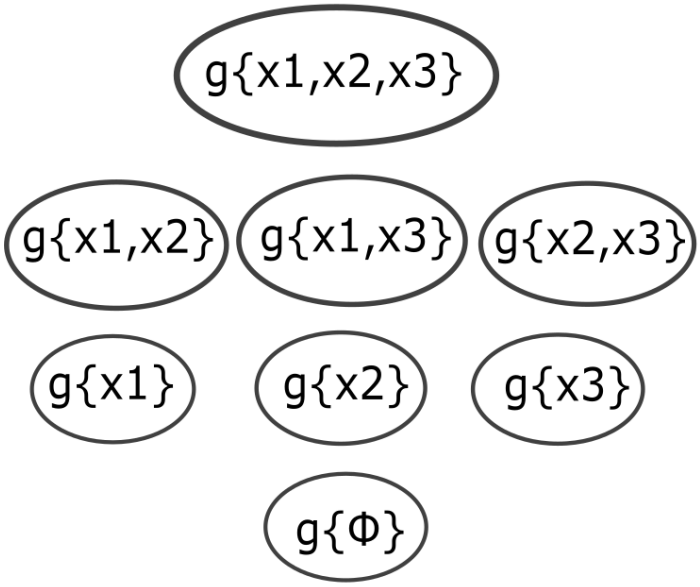


Figure 2.11 FM Lattice for three sources

For an infinite domain X there is an additional property to ensure continuity; However, it is not applicable in this work as X is finite and discrete. In the context of multi-source data fusion, $g(A)$ represents the weight or importance of subset A. The FM values of the singletons i.e. $g(x_i)$ are commonly called the densities. Three major approaches have been used to determine FMs:

1. *Expert-driven:* While the FMs can be defined by experts, with increasing lattice size (number of parameters), defining each parameter of the FM becomes practically not feasible, restricting this approach to a subset of applications with a limited number of sources [7]. For example, 30 coefficients need to be defined for 5 information sources.

2. *Algorithm-driven:* Several algorithm-driven measures including the Sugeno λ -measure, the S-decomposable measure, the possibility measure, the k-additive measure, the p-symmetric measure, the k-tolerant measure and the k-intolerant measure have been proposed in the literature [81–83, 7, 84]. Algorithm-driven methods compute the FM from the values of individual sources by leveraging the FM's mathematical constraints. These measures need the weights of the base sources to be defined in advance, i.e. these methods build the FMs from given source densities, constrained only by the two properties of the FMs. In contrast, some data-driven algorithmic measures such as the Meta-Measure and the Specificity use the criterion such as the agreement among subset of sources.

3. *Optimisation:* Various methods including evolutionary algorithms, gradient descent and quadratic programming have been used to generate FMs [7]. The FMs defined by these methods consider the entire lattice as independent weights, loosing the semantic meaning of the FM. Additionally, these methods optimise the FM *in respect to a pre-defined aggregation operator* and training data [85, 86, 23], focusing on minimising the training error and thus they do not optimise the independent meaning of a FM. For example, the DeFIMKL algorithm (defined in section 2.4.3) derives the FM using an optimisation method and the training data, w.r.t. to an aggregation operator defined in advance. Note that this approach, while offering the potential of powerful and concise fusion operators, is also most at risk of generating non-generic FMs, i.e. FMs which are not independently interpretable, but are tuned to drive the specific aggregation operator (eg. CFI) which they were trained for [7].

In the next sub-section two FMs are discussed, one of them being the recently introduced Uriz FM and the other being the most extensively used FM across the aggregation community i.e. the Sugeno λ-measure .

Sugeno λ-measure

Introduced by Sugeno in 1974, the Sugeno λ-measure (g_λ) is a parametric approach to generate the FM, which constructs the entire lattice just from the densities [87]. The Sugeno λ-measure has been the most commonly used algorithmic FM in the literature.

In addition to properties P1 and P2 defined in the section 2.4.1, the Sugeno λ-measure centres on the following property:

P3: $\forall\, A, B \in X, A \cap B = \phi$,

$$g(A \cup B) = g(A) + g(B) + \lambda g(A) g(B)$$

where $\lambda > -1$. This parameter λ can be uniquely obtained by solving the following polynomial equation:

$$1 + \lambda = \prod_{i=1}^{n} (1 + \lambda g_i)$$

where $g_i = g(x_i)$. Note, if $\lambda < 0$ then the measure is sub-additive, if $\lambda = 0$ then it is a probability measure and if $\lambda > 0$ then the measure is super-additive.

Uriz FM

The authors in [88] recently proposed a new method i.e. Uriz FM, which learns the FM from the classification accuracies of individual classifiers, similar to *A Priori* FM proposed this thesis (refer to chapter 5 and 6 fore more details) which also learns the FM from classification accuracies. The two methods are however different in the way FM is computed, more details in Chapters 5 and 6. The Uriz FM was also motivated to illustrate the improvement in ensemble classification performance by refining the fusion step of the ensemble (using FMs), defined as follows:

For all $A \subseteq X$ and number of classifier (or features) $N: \{1,..,n\}$, the Uriz's FM is composed of two steps. First, the uniform FM g_u is given by:

$$g_u(A) = \frac{|A|}{n}$$

The second step utilises the results of individual classifiers as well the results of each possible combination of classifiers, given by Acc_A, $\forall A \subseteq X$, defined as follows:

$$g(A) = g_u(A) + \frac{\tanh(Acc_A - MeanAcc_{|A|})}{2n}$$

where, $MeanAcc_{|A|}$ is the average of the results with the same cardinality and tanh is the hyperbolic tangent (for more details on the FM, refer to [88]).

The information quantified by these FMs are combined using the aggregation operators, defined in the next subsection.

2.4.2 Aggregation Operators

Aggregation operators are mathematical functions that combine the information from multiple sources [7, 23]. In the context of evidence fusion, aggregation operators extend the concept of weighted averages and are able to capture the interactions among the sources, delivering non-linear combinations. There are many aggregation operators in the literature [89], but in this thesis we focus on commonly used FIs and the recently introduced RAV, explained in the next subsections.

Weighted Average

Weighted average ensemble are models combining outputs by giving different weights to different inputs, usually based on importance. Let X be a set of k information sources. The

weighted average output $WA(x)$ for each evidence x_i (where $i = 1, ..., n$) is given by:

$$WA(x) = \sum_{i=1}^{k} w_i x_i, \tag{2.15}$$

where W_is are the set of weights (importance) for each corresponding instance such that $\sum w_i = 1$.

Fuzzy Integrals

Fuzzy Integrals (FIs) are often used as non-linear aggregation functions which combine information from multiple sources using the worth of each subset of sources (provided by a FM 'g') and the support of the question (the evidence 'h') [7, 23]. The two most commonly used FIs in the literature are the Choquet Fuzzy Integral (CFI) and the Sugeno Fuzzy Integral (SFI) [90], although in this thesis the CFI is the focus, defined as follows:

Choquet Fuzzy Integral: Let $h : X \to [0, \infty)$ be a real valued function that represents the evidence or support of a hypothesis. The discrete Choquet Fuzzy Integral (CFI) [23, 91, 7, 90] can be defined as:

$$\int_{CFI} h \circ g = CFI_g(h) = \sum_{i=1}^{n} h(x_{\pi(i)})[g(A_i) - g(A_{i-1})] \tag{2.16}$$

where π is a permutation of X such that $h(x_{\pi(1)}) \geq h(x_{\pi(2)}) \geq ... \geq h(x_{\pi(n)}), A_i = [x_{\pi(1)}, ..., x_{\pi(n)}]$ and $g(A_0) = 0$. More detail on the property of FIs and the CFI can be found in [89].

Recursive Average

The Recursive Average (RAV), an instance of Recursive Weighted Power Mean − introduced by Wagner et al. [7], is an aggregation operator that fuses information over a set of sources X defined with respect to a FM g, mathematically defined as follows:

$$RAV_p(X) = \begin{cases} h(X), & \text{if } |X| = 1. \\ \left(\frac{\sum g(B_j) * RAV_p(B_j)^p}{\sum g(B_j)}\right)^{1/p}, & \text{otherwise.} \end{cases} \tag{2.17}$$

where $|p| > 0, B_j = X \backslash x_j, x_j \in X$.

$\forall p$ the RAV for a set of sources is recursively defined as the weighted average of its sub sources, where the weight at each node is captured by a FM [7]. For some particular p values, RAV adopts specific averaging behaviour in a recursive manner, i.e. the recursive arithmetic

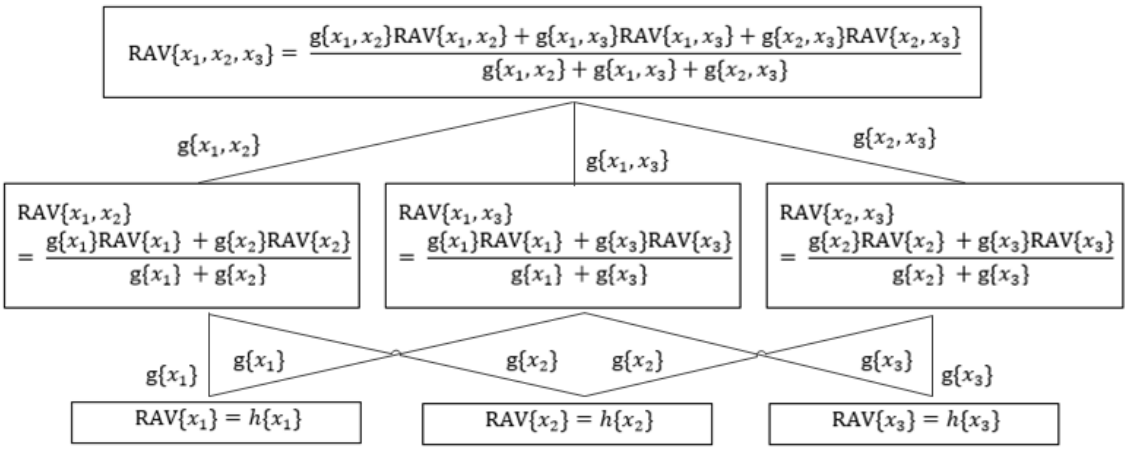


Figure 2.12 RAV for three sources (replicated from [7])

average for $p = 1$, the recursive harmonic average for $p = -1$, the recursive quadratic average for $p = 2$, and so on. For simplicity and space, we focus on solely on the RAV for $p = 1$ in this paper. Figure 2.12 shows the RAV computation for three sources with $p = 1$.

2.4.3 Aggregation operator based ensemble classifiers

Many traditional ensemble classification techniques exist in the literature which have been discussed in detail in the section 2.2. These ensemble methods have been very popular in the machine learning community due to their ability of producing more accurate results than individual classifiers [42, 55] in a very wide range of application areas [92]. Recently, another approach to obtain ensemble classification is the use of the aggregation operators defined with respect to a Fuzzy Measure (FM) [23, 91, 85, 86, 93]. In the previous studies [23, 94], the researchers concluded that aggregation operator based ensemble classifiers worked well in a number of applications such as multi-criteria decision making (MCDM) [95], forensic science [96], software defect prediction [97], brain computer interface (BCI) [98], computer vision [99, 100] and explosive hazard detection [101].

Based on the expected potential of non-linear aggregation offered by the aggregation operators (such as the CFI), its application to decision-level fusion in ensemble classifiers, i.e. to fuse multiple classifiers outputs towards one superior decision level output, has recently been explored. A key example of such a aggregation operator based ensemble classification method is the Decision-level Fuzzy Integral Multiple Kernel Learning (DeFIMKL) algorithm [23], which aggregates the outputs of kernel based classifiers through the use of the

CFI with respect to a FM learned through a regularised quadratic programming approach, explained in the next sub-section.

DeFIMKL

Let $f_k(\mathbf{x_i})$ be the normalised output generated by the kth classifier in an ensemble on a feature vector $\mathbf{x_i}$. The overall decision of the ensemble is computed by the Choquet Integral, where g encodes the relative worth of each classifier in the ensemble. Thus, the output of the ensemble with respect to the FM g on feature-vector $\mathbf{x_i}$ is produced by $f^g(\mathbf{x_i})$, mathematically described as follows,

$$f^g(\mathbf{x_i}) = \sum_{k=1}^{m} f_{\pi(k)}(\mathbf{x_i})[g(A_k) - g(A_{k-1})], \tag{2.18}$$

where $A_k = f_{\pi(1)}(\mathbf{x_i}), ..., f_{\pi(k)}(\mathbf{x_i})$, such that $f_{\pi(1)}(\mathbf{x_i}) \geq f_{\pi(2)}(\mathbf{x_i}) \geq ... \geq f_{\pi(m)}(\mathbf{x_i})$. It can be shown that (2.18) can be reformulated as

$$f^g(\mathbf{x_i}) = \sum_{k=1}^{m} [f_{\pi(k)}(\mathbf{x_i}) - f_{\pi(k+1)}(\mathbf{x_i})]g(A_k). \tag{2.19}$$

where $f_{\pi(m+1)} = 0$. Pinar et al. [23] proposed to learn FM g using a regularised sum of squared error (SSE) optimisation, described as follows,

$$E^2 = \sum_{i=1}^{n} (f^g(\mathbf{x_i}) - y_i)^2 + v(u), \tag{2.20}$$

where y_i is the class label for $\mathbf{x_i}$ and $v(u)$ is a regularisation function. Equation (5.2) can be further expanded as

$$E^2 = \sum_{i=1}^{n} (H_{x_i}^T * u - y_i)^2 + v(u), \tag{2.21}$$

where y_i is the actual class label for $\mathbf{x_i}$, u is lexicographically ordered FM g i.e. $u = (g\{x_1\}, g\{x_2\}, ..., g\{x_1 \cup x_2\}, g\{x_1 \cup x_3\}, ..., g\{x_1 \cup x_2 \cup ... \cup x_m\})$, and

$$H_{x_i} = \begin{pmatrix} . \\ f_{\pi(1)}(\mathbf{x_i}) - f_{\pi(2)}(\mathbf{x_i}) \\ . \\ . \\ 0 \\ . \\ . \\ f_{\pi(m)}(\mathbf{x_i}) - 0 \end{pmatrix}, \tag{2.22}$$

where H_{x_i} is of size $(2^m - 1)$ and contains all the difference terms $f_{\pi(k)}(\mathbf{x_i}) - f_{\pi(k+1)}(\mathbf{x_i})$ at the corresponding locations of A_k in u. We can fold out the square terms from (5.3), producing

$$\begin{aligned} E^2 &= \sum_{i=1}^{n} (u^T H_{x_i} H_{x_i}^T u - 2 y_i H_{x_i}^T u + y_i^2) + v(u) \\ &= (u^T D u + f^T u + \sum_{i=1}^{n} y_i^2) + v(u), \end{aligned} \tag{2.23}$$

where D and f are

$$D = \sum_{i=1}^{n} H_{x_i} H_{x_i}^T, \quad f = -\sum_{i=1}^{n} 2 y_i H_{x_i}$$

Equation (5.5) is a quadratic function and thus we can add the constraints on u such that it represents a FM, producing a constraint QP. We can add the monotonicity constraint on u according to the properties P1 and P2 as $Cu \leq 0$, such that

$$C = \begin{pmatrix} M_1^T \\ M_2^T \\ . \\ . \\ M_{n+1}^T \\ . \\ . \\ M_{m(2^{m-1}-1)}^T \end{pmatrix}, \tag{2.24}$$

where $M_1^T..M_{m(2^{m-1}-1)}^T$ are vectors representing monotonicity constraint such as the one used in this work i.e. $g\{x_1\} - g\{x_1 \cup x_2\} \leq 0$ (see [91] for more details on C). Thus, the full QP to learn FM u is

$$\min_u \ 0.5u^T\hat{D}u + f^T u + v(u), \quad Cu \leq 0, \quad (0,1)^T \leq u \leq 1, \tag{2.25}$$

where $\hat{D} = 2D$. We test the performance using ℓ_1 regularisation, i.e.

$$\min_u \ 0.5u^T\hat{D}u + f^T u + \lambda||u||_1, \tag{2.26}$$

where λ is the regularisation weight. The QPs at (5.6) and (2.26) provide a method to learn the FM u (i.e. g) from the training data. A new feature vector x', from a test set, can thus be classified using the following steps:

1. Compute the normalised decision value $f_k(x')$ for a classifier (SVM in this thesis),
2. Apply the CFI at equation 2.16 with respect to the learned FM g,
3. Compute the class label using $sign(f_k(x'))$.

The DeFIMKL is a key example of aggregation operator based ensemble classification approach. We will use this approach later in chapters 5 & 6 for comparison purposes (with the new FM proposed in this thesis). In this thesis, several applications are analysed to show the research gaps that persists, explained in the following sub-sections.

2.5 Ensemble Classification Applications

This section moves beyond different ensemble classifier to discuss the literature on how the ensemble classification algorithms and have been used in several applications, specifically within Breast Cancer, Heavy Good Vehicles, Face Recognition and standard datasets from UCI machine learning repositories.

2.5.1 Breast Cancer Biological Groups Stratification

Breast Cancer is one of the most common causes of cancer death in women, representing a very complex disease with varied molecular alterations [9, 8, 10]. It creates the particular difficulty of determining the suitable treatment option, among a number of complex treatment

choices. Therefore, classification of cancer patients into biological groups is essential for appropriate diagnosis to increase survival. In 2018 approximately 14 million new cancer cases and 9.6 million cancer related deaths were reported worldwide. Breast cancer is the second most common cancer with approximately 11.5% and fifth most common cause of cancer death [102, 103].

The most frequent symptom or early sign of breast cancer is a lump or mass, developed in the breast tissue [102]. Often, breast cancer is located in the lymph nodes under the armpits or above the collar bones. The most common types of breast cancer are: Ductal carcinoma in situ, Lobular carcinoma in situ, Invasive, Invasive ductal, Invasive lobular, Inflammatory, Pagets and some rare types. Apart from different types, breast cancer characterises a very diverse category of tumours in reference to genetic composition, physical composition and response to therapy. It creates a particular difficulty of determining the most suitable treatment option among a number of complex treatment choices available. To tackle this heterogeneity, methods have been developed to facilitate personalised breast cancer treatment and support clinical decision making by predicting the patients outcome.

Previously, a series of papers, largely published in the breast cancer domain have leveraged Computational Intelligence (CI) developments and tools, resulting in ground breaking advances such as the classification of cancer into newly identified classes - leading to improved treatment options. This development evolved in the past two decades is scattered among several works. However, the current literature on the use of CI to achieve this is fragmented, making further advances challenging [8, 10, 104–106]. The next sub-sections capture the developments in this domain so far, and simultaneously establish the research gaps. Subsequently, chapters 3 & 4 will address these research challenges in this important area by leveraging the strong potential of classification algorithms.

Division of patients into core classes

Eisen et al. [107] introduced hierarchical clustering combined with the visual study of dendrograms to analyse the relation among the gene expression. The work inspired the use of clustering techniques in the cancer profile with respect to genes data [108–110]. Perou et al. [109] used the hierarchical clustering algorithm on gene expression data identifying four molecular classes: luminal, HER2, basal and normal. A subsequent study further subdivided the luminal into three, namely A, B and C [111]. Sotiriou et al. [112] carried forward the research partitioning the basal group into two, and eliminating the normal group, thus obtaining a total of six groups. With advancements in technology, an alternative

approach developed is to use immunocytochemistry on formalin fixed paraffin embedded patient tumour sample. To address the stability issue of the classification from clustering techniques, Monti et al. [113] introduced consensus clustering, which considers the consensus across multiple clustering algorithms. Kellam et al. [114] refers to consensus clustering as Clusterfusion, which takes into account the results of different clustering algorithms and creates robust clusters depending on the consensus across the results of multiple clustering algorithms.

To assist prognosis in the context of breast cancer treatment, researchers have applied multiple protein biomarkers with relevance to breast cancer to investigate clinical importance of breast cancer classes [9, 114, 113, 105, 8, 66, 10]. Soria et al. 2010 built upon these works and analysed a breast cancer panel composed of 25 protein biomarkers, using the aforementioned clusterfusion approach [8]. The work aimed to develop a methodology to elucidate core and stable groups within this breast cancer dataset. The dataset incorporated the Immunohistochemical reactivity of 25 proteins with relevance to the breast cancer for 1,076 patients from the Nottingham tenovus primary breast carcinoma series, with primary operable stages (I, II and III). The work inspected five clustering techniques namely K-Means, PAM, Adaptive Resonance Theory (ART) and HCA. After selecting the clustering algorithms the next step was to determine the number of clusters. Therefore, clustering was followed by the computation of validity indices to validate the number of clusters and to tackle the issue of cluster stability. Six cluster validity indices were performed on the K-Means and the PAM, to evaluate the best number of clusters for the dataset. K-Means and PAM were chosen to run for 2 to 20 clusters for six indices: Calinski and Harabasz, Hartigan, Scott and Symons, Marriot, TraceW and Trace$W^{-1}B$. The next step was to perform the concordance between the clusters formed by the different clustering algorithms, assessed statistically by the Cohen's kappa coefficient, and visually via projecting the first and second components of the Principal Component Analysis (PCA). The last step was to characterise the core groups (classes) by Artificial Neural Network, Orthogonal Search Rule Extraction and boxplots. By establishing the consensus the work identified six novel cancer subtypes, and classified the patients among these core classes. The process identified prime biological classes: luminal, HER2 and basal (luminal and basal containing the remaining three distinct subclasses), shown in the breast cancer classification structure in Figure 2.13. The procedure classified 663 (61.6%) patients into one of the core classes while the remaining 413 (38.4%) patients were assigned to a mixed class i.e. 'not classified' (NC) group. Further analysis was required to minimise the mis-classification.

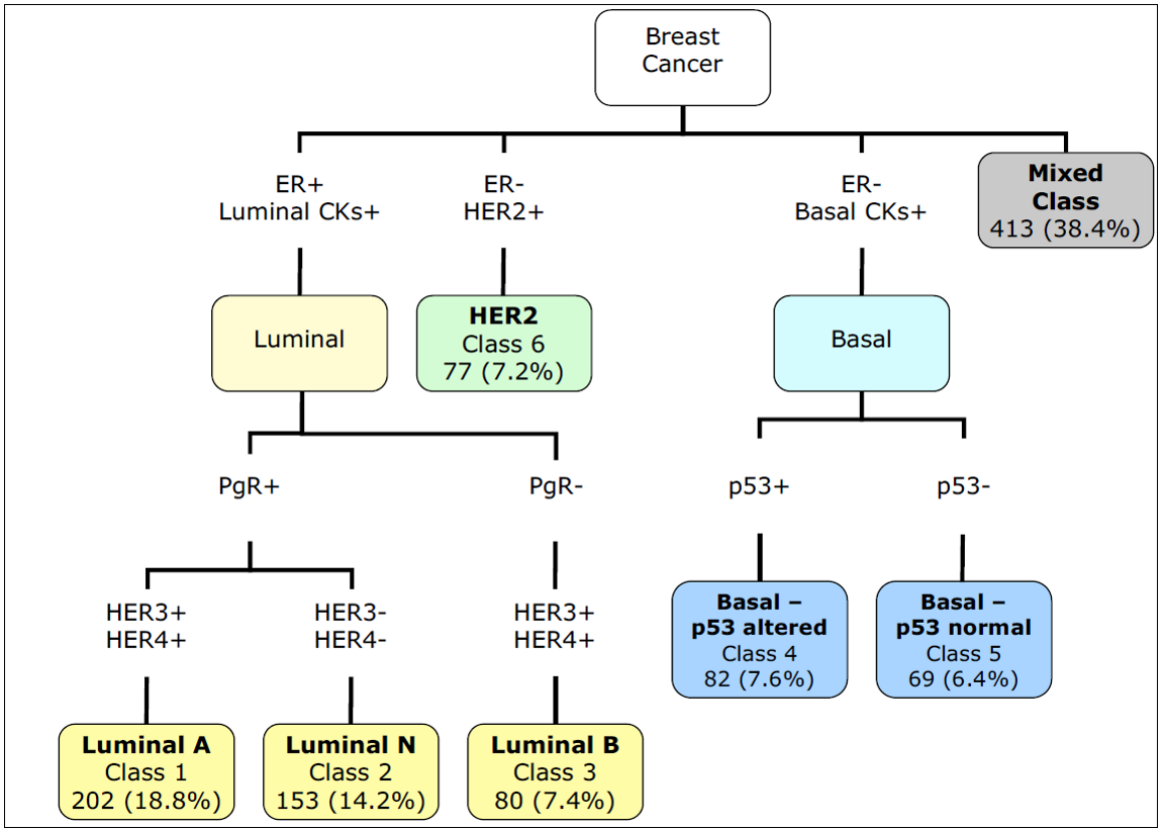


Figure 2.13 Breast Cancer Classification structure by Soria et al. [8].

Soria et al. [115] further verified the methodology developed in the previous work i.e. [8], to extract groups from a new dataset of breast cancer histone biomarkers which consisted of 347 patients. The dataset was provided by the Division of Molecular and Cellular Sciences in the School of Pharmacy at the University of Nottingham. The methodology identified 3 core classes, dividing 91.1% of the patients into one of the identified classes and leaving behind 31 patients, who ere assigned to the 'not classified' (NC) group.

Reduction of the biomarkers to an essential set

In a different set of study, Soria et al. 2008 [66] performed a comparison of three different classifiers for validating classification of breast cancer markers has been studied. The study was also performed on the Nottingham tenovus primary breast carcinoma series. The study was done on the 663 cases out of 1076 which were classified to one of the six groups. The classifiers used were Naive Bayes algorithm, C4.5 decision tree and Multilayer perceptron (MLP) algorithms. The tests have been done by taking 25 biomarkers and also by testing the classification algorithms on 10 'most important' biomarkers. The paper concludes that

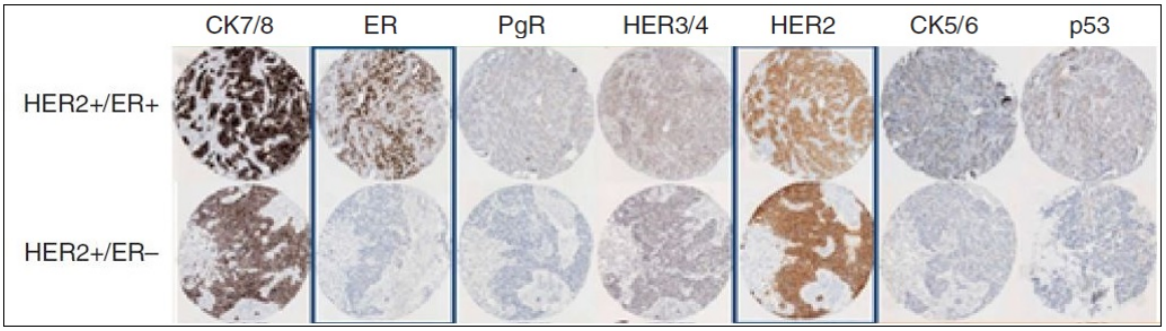


Figure 2.14 Subdivision of HER2+ in two subclasses HER2+/ER+ and HER2+/ER- based on immunohistochemical profiles by Green et al. [9].

when all the 25 markers were used for 663 cases, MLP gave the best result. The naive Bayes classifier and C4.5 returned similar results. When the biomarkers were reduced to 10 most important, MLP was still the best. The performance of Bayes improved and misclassified cases were reduced indicating the need to reduce the number of biomarkers.

In a subsequent work Green et al. 2013 [9] studied boxplots of the complete Nottingham tenovus primary breast carcinoma series dataset dataset and individual classes (described in Soria et al. [8]) and reduced the biomarkers on which the classes were derived, from 25 down to 14. The panel was reduced by omitting the biomarkers which had identical overall distributions across multiple classes. The study focused on the reduction of the number of biomarkers which were sufficient to obtain good classification. Using a Naive Bayesian supervised classification approach (as used in Soria et al. 2008), the set of biomarkers was further reduced to contain the ten 'most important' ones. The exhaustive search using the naive Bayes classification to choose the best combination of ten biomarkers out of 14 was done to also reduce the costs of clinical tests. This work also studied the class characterisation by analysing the immunohistochemical profiles of the patients. Class 6 (HER2) showed heterogeneity in the expression of the hormone receptor which led to division of HER2 into two subclasses HER2+/ER+ and HER2+/ER-, shown in Figure 2.14. The stable clusters validate the core classes (by Soria et al. [8]) and forms basis of further study (by Soria et al. 2013).

FuzzyQSBA classification rules

Rasmani et al. [116] introduced the implementation of replacing crisp weights in the Weighted Subsethood-based Algorithm (WSBA) by fuzzy quantifiers. The main difference of Quantifier Subsethood-based Algorithm (QSBA) compared to WSBA lies in the interpretation of the

inference between weights/quantifiers with the linguistic terms. In QSBA both the quantifiers and the linguistic terms are fuzzy sets. It offers us the flexibility to use t-norm operators to interpret the rule. The four datasets used are to test and compare the results: Iris-Plant, Wine Recognition, Glass Identification and Wisconsin Breast Cancer Dataset. The results show that classification accuracy of QSBA is as good as WSBA, and in Iris-Plant dataset QSBA outperforms WSBA.

The paper in [105] presents a rule simplification technique to extract linguistic rules from a data-driven quantifier-based fuzzy classification system. The dataset used for the comparative study was the Nottingham tenovus primary breast carcinoma series. The study was done on the 663 cases out of 1076 which were classified to one of the six groups. The paper shows that linguistic rulesets generated from a fuzzy technique have better advantages compared to the non-fuzzy in terms of 1) readability of the rulesets, 2) explicitly expressed characteristics of each rule and 3) providing information on the firing strength of each rule. The rulesets were consistent throughout the ten-fold cross validation tests.

The authors build upon the aforementioned research and use a similar methodology to Soria et al. 2010 and performs a consensus clustering to identify core classes on the Nottingham tenovus primary breast carcinoma series [117]. The work than makes use of rule simplification process (Rasmani et al. 2009) to construct a Fuzzy Quantifier Subsethood-based Algorithm (FuzzyQSBA). The algorithm results in a simple, comprehensible classification table for each of the six classes based on only ten biomarkers. The algorithm is constructed with aim to implement the resultant fuzzy rule table in a model-based classification system that can be used to determine the type (class) of cancer in new patients presenting with breast cancer.

Division of new patients into core classes by learning the pattern

Soria et al. [10] built upon the previous set of works ([8, 9, 105, 117]), and explored the use of fuzzy methodologies to generate a rule set which can classify a large number of patients into one of the identified classes. The work makes use of a data-driven fuzzy rule based system (FuzzyQSBA) for classifying the breast cancer patients. The FuzzyQSBA itself is similar to the algorithm developed in Rasmani et al. 2009 [105]. This latter study makes use of 663 classified patients reported in Soria et al. [8]. The algorithm developed in this work was applied to a well characterised dataset of 1,073 breast cancer patients from the same Nottingham tenovus primary breast carcinoma series. The patients were classified into one of the identified seven classes using fuzzy iteration method, similar to the algorithm developed

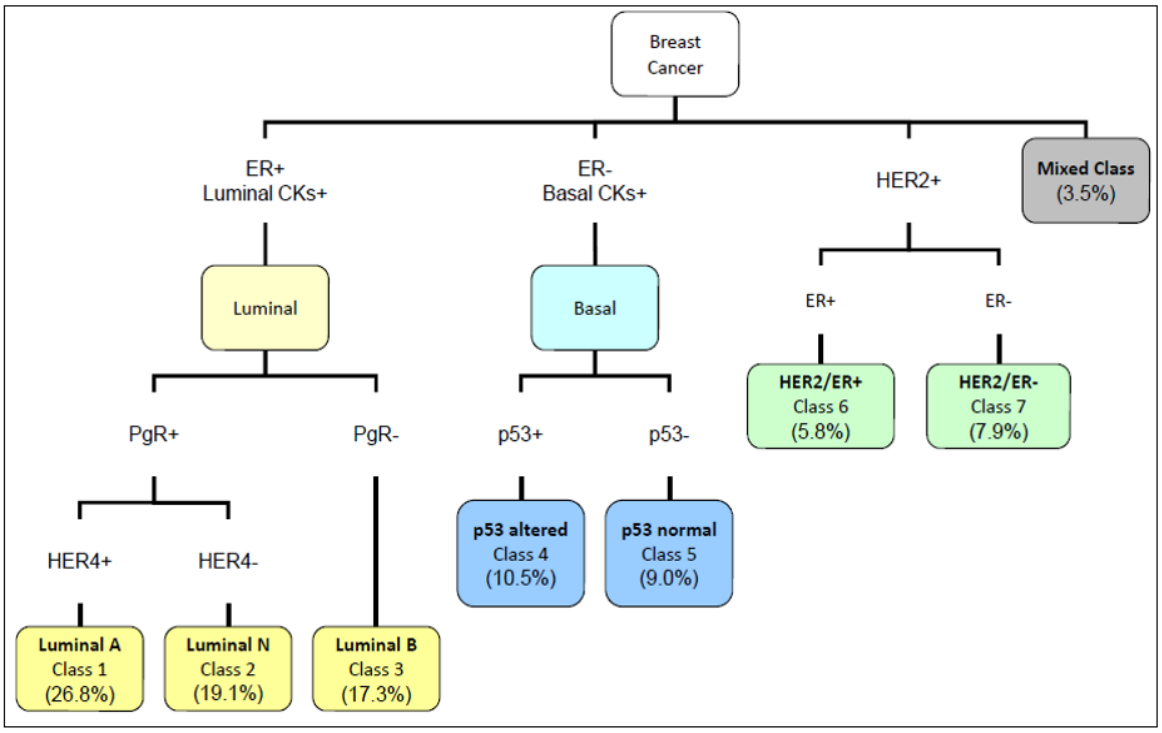


Figure 2.15 Breast Cancer Classification structure by Soria et al. [10].

in the work Green et al. 2013. The FuzzyQSBA algorithm learns the class-labelled breast cancer dataset by generating one classification rule per class. The algorithm assigns a class membership value to each patient for each of the seven classes. A novel class allotment algorithm is developed which assigns patients into one of the seven classes. In doing so, 38 out of 1,073 patients were 'Not Classified'. The class distribution is represented in Figure 2.15. The classification rules were validated on a new breast cancer dataset of 238 patients. Only 12 patients were assigned to 'Not Classified', giving an excellent classification outcome. In this thesis, we build upon the patients which were 'Not Classified', and propose to classify as many patients among one of the groups.

Application of the breast cancer biological group classification

The aim of this work by Green et al. [106] is to improve the current breast cancer prognostic methods by combining the well-established clinicopathological variables with biomarkers information. The methodology is developed on the 1,073 patients from the Nottingham tenovus primary breast carcinoma series, divided into two phases: 1) Identification of biological classes 2) Developing a NPI-like formulae, known as NPI+. The NPI is a prognostic index formula which is used as risk classifier in unselected cohorts of operable early stage

primary breast cancer patients. The NPI could not fully reveal the full survival outcome heterogeneity, so adding biological characteristics of breast cancer could significantly improve personalised medicine in breast cancer patients. The first phase was carried forward by Soria and colleagues in the work Soria et al. 2013, Soria et al. 2010, Soria et al. 2008, Green et al. 2013 and Rasmani et al. 2009. Using a consensus clustering based model the breast cancer patients were classified among seven core classes using 10 biomarkers. The second phase is the development of NPI+, providing improved treatment decision by: 1) combining biological features with already present parameters, 2) predict risk of disease recurrence and 3) improved prognostic analysis.

Green et al. [104] validated the capabilities of the breast cancer classification in prognosis using NPI+ on Edinburgh series breast cancer dataset. The dataset consists of 885 patients which were divided into seven classes using the same set of ten biomarkers panels as used in the previous works by Soria and colleagues. The NPI+ biological classes generated were similar to the ones developed in the aforementioned work [106], which points to the robustness of the NPI+ formulae. Note: This Edinburgh breast cancer is also employed in Chapter 3 to validate the proposed framework.

2.5.2 Heavy Goods Vehicle Driving Stereotypes

Although driving behaviour has been largely studied amongst private motor vehicles drivers, the literature addressing heavy goods vehicle (HGV) drivers is scarce. Identifying the existing groups of driving stereotypes and their proportions enables researchers, companies and policy makers to establish group-specific strategies to improve safety and economy. In addition, insights into driving styles can assist predicting drivers' reactions and therefore enable the modelling of interactions between vehicles and the possible obstacles encountered on a journey. Consequently, there are also contributions to the research and development of autonomous vehicles and smart roads. Thus, in this thesis our interest lies in investigating driving behaviour within the HGV community in the United Kingdom (UK).

Research in driving behaviour, vehicle monitoring, driving style prediction and driver modelling are on the rise, as there is a global demand for intelligent solutions to improve driving economy and road safety. The challenge lies in determining solutions capable of addressing social, financial and geographical requirements within a problem context. To achieve this, the investigation of actions and policies for education, law enforcement and improvements in infrastructure are of essence. Equally important is the understanding and prediction of the drivers responses to these interventions. This review is focused on the work

carried out to understand, model and predict driver behaviour. Current literature targeting solely HGV drivers is scarce. In addition, there is not much literature targeting UK HGV drivers. We therefore include in our review studies concerning the analysis of specific driving behaviour for both commercial and non-commercial drivers, obtained through the analysis of telematics, GPS and other mobile devices data. Studies on general driving style for both small vehicles and HGV drivers obtained through self-assessment questionnaires [118, 119] is far more common. However, due to the differences in the nature of the data collected, we do not include this research in our review. Our objective is therefore to identify the research gaps in telematics data findings and to establish a baseline for comparison with the methodology proposed in chapters 3.

A relevant work in the area of driving behaviour was conducted by Constantinescu *et al.* [12], in which six driving profiles were identified from telematics data. Data was obtained in a five-day experiment, involving 23 drivers and two control drivers with up to nine journeys each. Controls were employed to introduce extreme behaviours (very aggressive and slow, non-aggressive, economical). 200 journeys were analysed, and the features considered were the percentage of time above the speed limit (60kph), the mean and standard deviation (std.) of speed, the mean and std. of acceleration, the mean and std. of braking and the total energy required to increase the speed. Hierarchical clustering (HC) identified 6 clusters and Principal Component Analysis (PCA) assisted in their interpretation and in the detection of four factors determining behaviour: aggressiveness, speed, acceleration and braking. For each factor, the authors identified the range values and how drivers within the clusters behaved. Ranges were converted into labels characterising drivers' stereotypes, (Table 2.2). Although the findings were an important contribution to understand driving behaviour, this work is limited to the number of drivers and journeys investigated. More data needs to be collected to further validate the clusters detected. In addition, control subjects forced the analysis to establish pre-defined clusters with extreme behaviour. These artificial clusters also need to be further contrasted with more data instances to assess their likelihood of occurrence in real-world scenarios.

Kalsoom and Halim [120] applied K-Means (KM) and HC to a dataset obtained via simulation of 30 drivers. Data collected included the number of left and right turns, left and right indicators, brakes, horns and gear change with speed. Three driving styles were established, slow, normal and fast drivers. This work is also limited to amount of drivers studied. Less clusters of driver behaviour were detected, which might suggest that either the simulation experiments or the cluster analysis did not capture all patterns of behaviour found in Constantinescu *et al.* [12].

Table 2.2 Driving behaviour clusters from Constantinescu *et al.* [12]

Group	Aggressiveness	Speed	Acceleration	Braking
1	Moderately low	Low-Moderate	Moderate	Smooth-Moderate
2	Very low	Low-Moderate	Low-Moderate	Smooth-Moderate
3	Moderately high	Moderate	Moderate	Sudden
4	Neutral	Moderate	High	Moderate
5	Neutral	Moderate-High	Low-Moderate	Moderate-Sudden
6	High	High	High	Sudden

Castignani *et al.* [121] collected data from motion sensors and GPS, which is later analysed to determine driving events. Fuzzy logic was employed to detect harsh braking, harsh acceleration, over speeding and aggressive steering. The events obtained were subsequently merged with time and weather information to determine their risk level and to score the driver. Experiments employed the same vehicle over a predefined path in Luxembourg. Journeys occurred at daytime, with different weather and predefined behaviour. For the first lap the driver was supposed to drive calmly, observing the speed limit and avoiding abrupt manoeuvres; in the second lap the drivers behaved aggressively. Results showed consistency in the number of incidents obtained during calm and aggressive driving. One of the limitations of the study is that it tested the efficiency of the fuzzy system over predefined driving behaviour. In real-world scenarios, however, variation of driving styles and unplanned actions are likely to occur.

Halim *et al.* [122] employed KM, fuzzy c-means and Model-Based Clustering (MBC) to determine four driving profiles from 50 drivers. Data was acquired with a hardware consisting of vehicle functionalities and a simulator software with a virtual driving environment. Each subject drove the car in three scenarios (high, average and low traffic) for 15 minutes per scenario. The variables considered were number of left indicator, left turns, right indicator, right turns, brake use, horn use, reverse gear use, average gear, maximum gear, average speed, maximum speed and gender. KM results were compared with those from fuzzy c-means and MBC to establish the optimal number of clusters. PCA determined the best features (average and maximum speed, brakes and horns were selected). Profiles 1 and 4 comprised of cautious drivers with high average speed but appropriate number of brakes. Profile 2 characterised slow and sluggish drivers with high frequency of braking and indicators. The third profile included safe drivers with moderate speed, number of use of indicators and brakes.

Ellison *et al.* [123] proposed a risk index framework for scoring drivers based on their behaviour, personality, perceptions, demographics, vehicle characteristics, weather conditions and time of the day to assess the risk of crashes. 8 million GPS data observations with 1Hz

frequency from 106 drivers in Sydney over several weeks were considered. Results indicated that over 90% of drivers exhibit more variability in speeding, acceleration and braking behaviour between different road environments when compared to the same road. The author's results suggest the potential for using more dis-aggregate data but also the necessity to control for temporal and spatial factors when studying driver behaviour. These conclusions reiterate the importance of our study, where a larger data sample should provide us with better insights into HGV driving behaviour.

Similarly, Saiprasert *et al.* [124] calculated a driver's safety index based on their driving events and categorise the driver into four safety level profiles (very safe, safe, aggressive, very aggressive). The calculation is based on an equation in which the events (harsh acceleration, harsh braking, harsh turning, sudden lane changing and over speed events) are weighted according to the characteristics of the road. Data collection was carried out with 20 drivers in two of Thailand's main highways, with 30 journeys per route. Results showed that the majority of the drivers are either safe or aggressive for both routes.

From the literature it is possible to identify the clear gap in the current research, that there is very little literature regarding driver profiling based on telematics data for the UK. Our thesis therefore aims at contributing in filling these research gaps, more details in chapter 4 section 4.2. It is important to notice that in addition to the aforementioned gap, this thesis is also able to address two more research gaps: (i) limited number of drivers have been investigated in the previous studies, and (ii) most experiments have been conducted within a small number of routes.

2.5.3 Aggregation Operators in Face Recognition Applications

Due to the recent advance in deep learning, Convolution Neural Networks (CNNs) have been successfully applied to various deep facial feature extraction, i.e. facial recognition [125, 126], facial shape [127, 128], spatial-temporal domain [129, 130]. Since this technique allows task-specific information to be extracted, CNNs normally yield better results than the traditional hand-craft features for face recognition tasks. In [131, 132], several CNN models were trained for extracting features from facial regions, which were then fused by concatenating representations yielded from all models. Compared to using holistic face images, region-based face recognition are more robust to pose variations and occlusion [133], and thus usually achieved superior results. Motivated by this, this thesis employs CNN as a base classifier for classifier-driven FM learning.

Aggregation operator based recognition tasks have received a lot of attention from the fuzzy community and a number of works have been presented since then. Tahani et al. [134] presented early works on the use of FI for facial image information fusion. Recently, Karczmarek et al. used the Sugeno λ-measure to aggregate salient facial regions and concluded that FMs owing to their monotonicity, leads to a better fusion model (in their case classification) [135]. The research group further confirmed the results in a subsequent study [136], by analysing a family of 25 t-norm operations to replace dot product in the CFI on multiple face recognition datasets. In another set of works, Chen et al. [137] presented a Multi-Region Ensemble Convolutional Neural Network (MRE-CNN), which uses both local face regions and the whole face for age prediction. Similarly, Fan et al. [138] used the MRE-CNN for facial image recognition. These work present the idea that considering local face regions will improve the overall performance.

2.5.4 UCI machine learning repositories

The UCI Machine Learning Repository is a collection of datasets, domain theories, and data generators that are used for the empirical analysis of CI algorithms. It has been widely used by students, educators, and researchers all over the world. These datasets are of high-quality and labelled, useful for all types of machine learning algorithms (specially supervised and semi-supervised). For demonstration, these datasets are used for the validation of the methodologies proposed in this thesis (more details in chapters 4, 5 & 6).

2.6 Summary

This chapter provides an overview of the existing literature that have been used in this thesis. In particular, several ensemble classification methods (and individual classifiers) have been extensively reviewed in this chapter. The strengths of these models were highlighted in addition to the key challenges faced by them. In additional to traditional classification approaches, unsupervised algorithms (i.e. clustering) based classification have also been reviewed.

In many case studies, aggregation operator based methods have obtained superior performances. A review has therefore also been made on aggregation operator based ensemble classification approaches, including their building blocks i.e. aggregation functions. Lastly, we also provided a general background for the case studies that are used throughout the

studies. These case studies allow us to evaluate the theoretical frameworks that we will propose in this thesis.

With background materials reviewed, in the upcoming chapters, we will investigate the mechanisms to improve research gaps in the ensemble classification.

Chapter 3

Combining Clustering and Classification to Enhance Classification of Unlabelled Data

One key challenge within unlabelled datasets is the classification of data points which are difficult to label. Although ensemble clustering approaches are able to elucidate characteristic groups, a number of data points are left behind as unclassified. Therefore, this chapter builds upon the research gap and introduces a novel and robust two-stage classification framework which employs a combination of both ensemble clustering and ensemble classification to improve the overall classification. Here 'robust' refers to flexible framework which can be updated depending upon application i.e. the algorithms can be modified according to the application and dataset.

Section 3.1 discusses the drawbacks of current clustering ensemble algorithms to elucidate the characteristic groups from unlabelled datasets, using Breast Cancer as a case study. Section 3.2 leverages a case study on Breast Cancer to demonstrate the need for an enhanced classification framework, introducing a new ensemble classification stage after the ensemble clustering stage for classification in an unlabelled dataset, therefore addressing first objective of the thesis. Subsequently, in Section 3.3 the resulting step-by-step classification framework is presented. coupling ensemble clustering with ensemble classification for the identification of core groups, data distribution in them and improvement in final classification results by targeting the unclassified data. Lastly, Section 3.4 summarises the contribution of this chapter to this thesis.

3.1 Background and Motivation

Determining the number of groups in a dataset with no class labels is a very challenging task [75]. One such problem is relevant in assisting breast cancer prognosis by determining the most suitable treatment options among a number of complex treatment choices available [8, 10, 106]. Recent studies have employed computational intelligence methods to facilitate personalised breast cancer treatment (by fully understanding the characteristics of the disease for each individual patient) and support clinical decision making by predicting the patient's treatment outcome. [8–10, 66, 104–106, 115]. In particular, clustering algorithms have addressed this problem by identifying treatment classes and dividing patients among them, a detailed discussion is presented in section 2.5.1 of chapter 2.

Previously, Eisen et al. [107] combined hierarchical clustering with the visual study of the dendrograms to determine the association between the gene expressions. Using clustering methods on breast cancer profiles, Perou et al. [109] recognised four distinct molecular classes of breast cancer based upon gene expression data: *HER2*, *basal*, *luminal* and *normal*. Extending the work, Sorlie et al. [111] subdivided the Luminal group into three: *A*, *B*, and *C*. Being uncertain of Luminal C, in a subsequent work [112] they divided basal into two and discarded the normal group and Luminal C, thus attaining an aggregate of six breast cancer groups.

Another problem lies in algorithm selection, as different algorithms produce different results. Therefore, contrasting and combining the results of multiple clustering algorithms is essential to test the stability of the clusters formed and of the proposed classification. Kellam et al. [114] used this idea and introduced 'Clusterfusion', which takes into account the results of all the different clustering algorithms present in the ensemble. The idea behind consensus clustering is to provide a level of confidence to the final grouping and simultaneously increase the robustness of sampling variability of the clusters formed. Note, if the robustness in sampling increases, the confidence of the overall classification structure also increases. Monti et al. [113] used consensus clustering in combination with re-sampling techniques to identify biological groups. Similarly, Chen et al. [139] introduced an ensemble clustering method to predict survival rate of cancer patients when working with discrete features. The algorithm used Partitioning Around Methods recursively to construct a dissimilarity matrix to be used by Hierarchical Clustering Algorithm to obtain final distribution.

In the past decade, Soria et al. [66, 8, 10] used the clusterfusion approach on data from immunohistochemistry on formalin-fixed paraffin embedded patient tumour samples, an alternative approach to gene expression profiling, for better identification of breast cancer

biological groups. The patients from Nottingham Tenovus Primary Breast Carcinoma Series were divided among seven breast cancer groups. Although the researchers have employed multiple relevant protein biomarker to large numbers of cases to elucidate breast cancer biological groups using ensemble clustering methods [8–10, 109, 140], the methodology left behind a number of unclassified patients, similar to previous works. Therefore, a framework still needs to be developed which can assign as many unclustered patients to one of the identified groups in order to improve final classification results.

This chapter builds upon the research gap and puts forward a novel two staged classification framework, which employs an additional ensemble classification layer after the ensemble clustering layer. To develop this new classification framework and ensure its systematic evaluation and suitability to address real-world challenges, we have developed the framework in respect to a Breast Cancer case study. Moreover, in addition to demonstrating the need for a new classification framework, this case study also addresses objective 1 and 2 of the thesis, as detailed in the following section.

3.2 A case study illustrating challenges in current classification frameworks

This section of the chapter presents a case study to motivate and illustrate the need to establish a clear step-by-step framework for classification of data points which are difficult to classify across the unlabelled datasets (note that further evaluation on the application of this framework on different case studies, including a different breast cancer study, is presented in chapter 4). In a seminal work by Soria and colleagues [8–10, 66, 105], analysis of immunohistochemical data (protein biomarkers) in Nottingham breast cancer dataset has led to formation of seven clinical groups (for more detail, refer to section 2.5.1). The researchers identified these groups by establishing a fuzzy rule based classification algorithm (FuzzyQSBA) and a novel class assignment method. Although this method established robust biological groups, 38 out of 1,073 patients were 'Not Classified' to any of the biological group. With the aim to establish a classification framework (which can classify a higher number of data into one of the identified groups), in this section we examine the proposed solution of using an additional '*ensemble classification*' layer. The complete set of steps described are represented in a flowchart diagram, shown in Figure 3.1.

To improve classification, we propose the use of ensemble classifiers to learn the classification rules by training on the well-defined biological group (from the clustering step) and

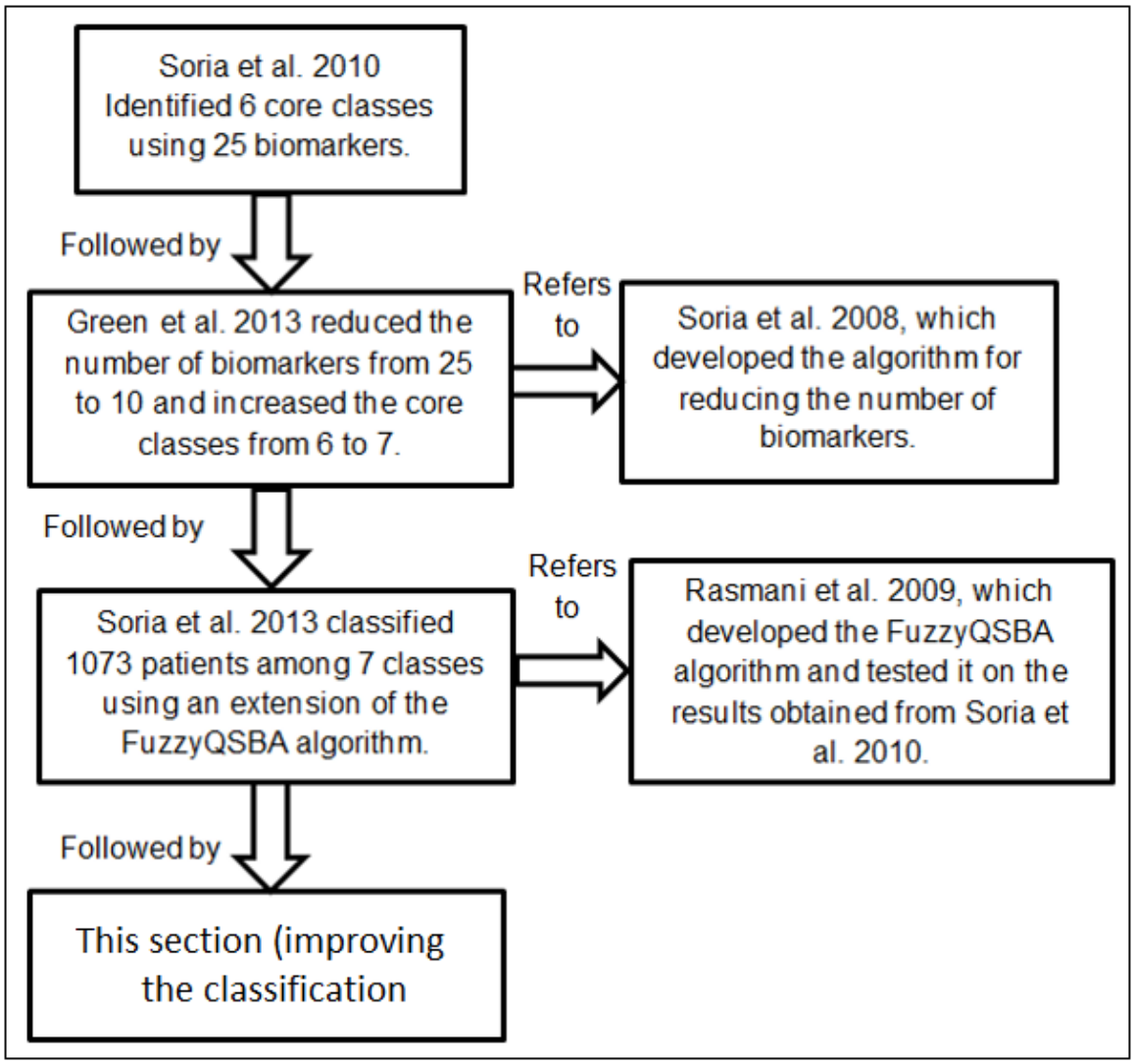


Figure 3.1 Flowchart of the research on Nottingham Breast Cancer Dataset

subsequently assign as many previously 'Not Classified' breast cancer patients to one of the specific group. Two specific datasets are used as part of this study, the first being the original breast cancer dataset and the second called 'fuzzy membership values dataset' produced as a result of the FuzzyQSBA algorithm mentioned above [10]. As the ground truth data does not exist for this type of cancer-based datasets, the resulting classification results are validated by employing a series of clustering techniques.

3.2.1 FuzzyQSBA Algorithm

The FuzzyQSBA algorithm [105] is a fuzzy rule induction method based on fuzzy subsethood values i.e. representing the degree to which a fuzzy set is a subset of another fuzzy set. The aim of using this technique is to replace crisp weights in the Weighted Subsethood-Based Algorithm (WSBA) [10] by fuzzy quantifiers. The main difference of Quantifier Subsethood-Based Algorithm (QSBA) compared to WSBA lies in the interpretation of the inference between weights/quantifiers with the linguistic terms. In QSBA both the quantifiers and the linguistic terms are fuzzy sets. It offers us the flexibility to use t-norm operators to interpret the rule.

The FuzzyQSBA algorithm consists mainly of these four steps: 1) Divide the dataset into subgroups. 2) Calculate fuzzy subsethood value. 3) Calculate relative weights. 4) Create fuzzy rules. Examples of fuzzy rules are as follows:

> WSBA - “IF A is (A1 OR 0.09A2) AND B is (B2 OR 0.2B3) AND C is (C1) THEN Output is D".
>
> FuzzyQSBA - “IF A is ((almost all)A1 OR (a little)A2) and B is ((almost all)B2 OR (almost a quarter of)B3) AND C is ((almost all)C1) THEN Output is D".

The above steps generate the likelihood of membership of each patient in the biological groups. After obtaining these membership values, the class assignment algorithm was run. For class assignment, the two highest membership values were chosen. If the difference between the top two values was greater than a certain threshold, then the patient was assigned to the class with maximum membership. If the difference was below the threshold and they both belong to same class family then the class with maximum values was chosen. If these conditions were not met, the patient was not classified.

3.2.2 Dataset Description

Nottingham Breast Cancer Dataset

The Nottingham Breast Cancer Dataset consists of 1,073 patients from the Nottingham Tenovus Primary Breast Carcinoma Series with operable stages I, II and III. Immunohistochemical reactivity [141] for 25 proteins was determined using tissue microarray (tumour samples). Among the available information, we consider the same ten important protein markers suggested by Green et al. [9] and run our test on them. The same reduced set of ten

					Fuzzy Membership values							
Patient	ER	PgR	...	MUC1	1	2	3	4	5	6	7	Class
1625	280	0	...	235	0.20	0.15	0.90	0.10	0.10	0.55	0.20	3
2879	0	0	...	130	0.10	0.15	0.20	0.65	0.20	0.10	0.10	4
3932	200	295	...	200	0.85	0.10	0.20	0.15	0.20	0.10	0.05	1

Figure 3.2 Tuples of data set produced after applying the FuzzyQSBA algorithm.

protein biomarkers has been used by Soria et al. [10] to classify patients in seven clinical classes using a data driven FuzzyQSBA algorithm.

Fuzzy membership value dataset

Fuzzy membership values, as shown in Figure 3.2, are the membership values obtained after implementing FuzzyQSBA on the ten protein biomarkers, numbered from 1 to 7, and form what we call the 'fuzzy membership value dataset'. The column Class represents one of the seven identified classes in which the patient is classified. ER, PgR, ..., MUC1 represents the ten protein biomarkers from the Breast Cancer Dataset on which the FuzzyQSBA algorithm was applied. The FuzzyQSBA algorithm [10, 105] uses continuous fuzzy quantifiers to create the rule-set, which assigns a class membership value to each patient for each of the seven classes. The advantage of using a fuzzy rule based model is its ability to represent the information in the form of if-then rules. This mechanism of representation is human understandable and combines several rules to generate global behaviour of the system.

3.2.3 Experiment

Ensemble Classification Model

In this case study a unanimous voting ensemble classifier is employed, which generates the final outcome by taking a consensus of results across all the classifiers. Unanimous voting is chosen as it provides robustness to the result by assigning a high level of confidence to the classification.

Experiment steps

Two algorithms namely: Artificial Neural Network (ANN) and Nearest Neighbour, were used for classification of the 'Not Classified' patients. The underlying strategy is as follows: 1) to classify 'Not Classified' patients in one of the identified classes, 2) to increase accuracy by considering multiple classification methodologies. To establish an accurate classification, the

algorithms were run on both datasets: the breast cancer dataset consisting of protein markers and the fuzzy membership values of each patient for the seven classes. The classification results were validated by applying three clustering techniques namely K-means, Hierarchical Clustering and K-Medoids. The steps followed were:

1. First, the ANN classification algorithm was run on the 1,073-patients breast cancer dataset, where 1,035 were classified in one of the seven classes. The ten protein biomarkers data of 1,035 patients were used to train the 3-layer neural network. The network was tested with multiple numbers of hidden nodes, and 20 was identified to be the optimal number. After the network was trained by updating its weight, the updated network was tested on the 38 'Not Classified' patients. Similar steps were adopted with the Fuzzy membership value dataset. The labelled data of seven class membership values for the 1,035 patients was used to train the neural network. After the network learned and updated its weight the updated network was tested on the 38 'Not Classified' patients. The classification outcomes of both the datasets are explained in the 'Results after Classification' section of the Results section.

2. The Nearest Neighbour algorithm calculates the distance of 38 'Not Classified' with all the labelled 1,035 patients. It assigns the class to 38 patients based on minimum distance. The same strategy was used in both the breast cancer dataset and the fuzzy membership values dataset. In the breast cancer dataset the distance was calculated among ten protein biomarkers, and in the fuzzy membership values dataset the distance was calculated among the membership values. The classification outcome for both the datasets is explained in the 'Results after Classification' segment of the Results section.

3. The classification results were verified by changing the parameters of the classifiers, followed by the consensus between the classification algorithms.

4. K-means clustering algorithm was run on both the breast cancer dataset and the Fuzzy Membership value dataset. Results of the K-Means clustering on the Fuzzy membership value dataset are explained in the 'Results after Clustering' segment of the Results section but the results of the K-means applied to the ten protein biomarkers of the breast cancer dataset are not reported as they were not informative and therefore dropped from further analysis.

5. Hierarchical and K-medoids clustering algorithms were run on both the breast cancer dataset and the Fuzzy Membership value dataset. Outcome of both the algorithms on

both the datasets are explained in the 'Results after Clustering' section of the Results section.

6. Consensus between the results of the different clustering algorithms was studied, and compared with the classification outcomes.

3.2.4 Results

Results after Classification

Classifusion works by generating a list which checks the output class of an instance for each algorithm. If the class is the same throughout, then the instance is assigned to the consensus class, otherwise the instance is considered to be misclassified. Table 3.1 shows the class labels for classifiers in columns (1) and (3) , for clustering algorithms in columns (2) and (4), and consensus of all in the last column 'Among all 4' respectively. Abbreviations of the table headers are as follows:

'Ann original': ANN algorithm applied to the 10 protein biomarkers data with 20 hidden layer nodes and 20,000 iterations of weight updating. 'Ann mem': ANN algorithm applied on the fuzzy membership values dataset. 'Nearest neigh original': nearest neighbour algorithm applied to the 10 protein biomarkers data. 'Nearest neigh mem': nearest neighbour algorithm applied on the fuzzy membership values dataset. '1 & 3': consensus among the ANN and nearest neighbour algorithm applied to the 10 protein biomarkers values. '2 & 4': consensus among the ANN and nearest neighbour algorithm applied to the fuzzy membership values. 'Among all 4': consensus among the ANN and nearest neighbour algorithm for both fuzzy membership values and 10 protein biomarkers values datasets.

Table 3.1 shows the classification results for ten randomly selected patients out of 38. Out of 38 patients, 31 had consensus results in case of '2 & 4' i.e. by combining the classification results classification of Fuzzy membership values dataset. In case of consensus of classification of breast cancer dataset, i.e. '1 & 3', the number reduced to 24. The consensus of classification for both the datasets ('Among all 4') further reduced down to 18 out of 38. It can be noted that the fuzzy membership values dataset generated better results. One explanation for its better results and might be that the fuzzy membership values dataset is pre-processed, i.e. the dataset is the result of FuzzyQSBA algorithm.

A number of observations can be drawn after the classification process: firstly, when the algorithms were run on the membership values dataset patients were either classified into classes corresponding to the highest membership value or to the second highest membership

Patient No.	Ann Original (1)	Ann mem (2)	Nearest neigh original (3)	Nearest neigh mem (4)	1 & 3	2 & 4	Among all 4
2	1	1	1	1	1	1	1
3	4	4	4	4	4	4	4
7	5	5	5	5	5	5	5
10	5	5	5	5	5	5	5
14	4	3	4	4	4	0	0
24	5	5	5	5	5	5	5
25	4	4	4	4	4	4	4
32	5	5	5	5	5	5	5
35	4	4	4	4	4	4	4
38	4	4	4	4	4	4	4

Table 3.1 Results after classification

value, but no patient was classified to class 6 or class 7. The membership values for class 6 and class 7 were a constant value for all the 38 patients. In addition, when the algorithms were run on the ten protein biomarkers similar results to those of the membership values were obtained, i.e. no patient was classified to class 6 or class 7. The reason for this was that the values of HER2 in the original 1,073 data were either 0 or 300. If the HER2 values were 300 then the patient belonged to either class 6 or class 7. If the value was 0 then the patient belonged to one of the remaining five classes or was misclassified. For all the 38 patients under investigation in this work, the values of HER2 were 0. So the patient would never belong to class 6 or class 7. We decided to omit class 6 and class 7 in further clustering analyses of 'Not Classified' breast cancer patients for the reasons explained above.

Results after Clustering

Individual results (for ten randomly selected patients) after each clustering algorithm along with consensus across all the methods are reported in Table 3.2. Abbreviation of the table headers are as follows:

'K-medoids orig': K-Medoids clustering algorithm applied on the ten original protein biomarkers. 'K-medoids mem': K-Medoids clustering algorithm applied on the fuzzy membership values. 'Hierarchical clustering orig': Hierarchical clustering algorithm applied on the ten protein biomarkers. 'Hierarchical clustering mem': Hierarchical clustering algorithm applied on the fuzzy membership values. 'K-means mem': K-Means clustering algorithm applied on the fuzzy membership values.

As reported in the classification discussion above, we eliminated the possibility of cluster 6 and cluster 7, and reduced the number of clusters in which the patients could be grouped

Patient No.	K-medoids orig (1)	K-medoids mem (2)	Hierarchical clustering orig(3)	Hierarchical clustering mem (4)	K-means mem (5)	Among all 5 (1&2&3&4&5)
2	1	1	1	1	1	1
7	5	5	5	5	5	5
8	4	4	4	4	4	4
14	2	4	2	4	4	0
16	1	1	1	1	1	1
21	4	5	4	5	5	0
22	3	3	3	3	3	3
26	4	5	4	4	5	0
35	4	4	4	4	4	4
38	4	4	4	4	4	4

Table 3.2 Results after clustering

to five. Columns (2) to (6) of Table 3.2 represent the outcome of K-medoids, Hierarchical clustering and K-means clustering applied on the breast cancer dataset and on the fuzzy membership value datasets. Out of 38 patients, 20 had consensus results across all the five methods. The classification result classified 18 out of 38 patients in one of the seven classes. Out of 20 patients from clustering, 18 patients are exactly the same as those from the classification results. The consensus of clustering and classification results therefore classifies 18 patients into one of the identified classes.

K-medoids and Hierarchical clustering produced the same results with respect to each other by generating similar clusters. They classified the same 20 patients in one of the identified classes for both the datasets, and performed slightly better than the classification by assigning more patients into one of the classes. The results obtained by applying K-means on the breast cancer dataset were not informative and were dropped from further investigation. But, the results obtained on the fuzzy membership values dataset were similar to those produced by K-medoids and Hierarchical clustering i.e. they classified the same 20 patients into their respective classes.

K-means clustering algorithm performed poorly on the breast cancer dataset, but it produced good results in case of the fuzzy membership values dataset. This difference could be due to the fact that the fuzzy membership values are processed values and have been obtained after applying FuzzyQSBA on original breast cancer dataset.

3.2.5 Discussion

After applying clustering and classification techniques on the 38 ‘Not Classified’ patients from the Nottingham breast cancer series, 18 of the patients can now be assigned to one of the seven classes. These patients show strong affinity with the original classes 4 and 5.

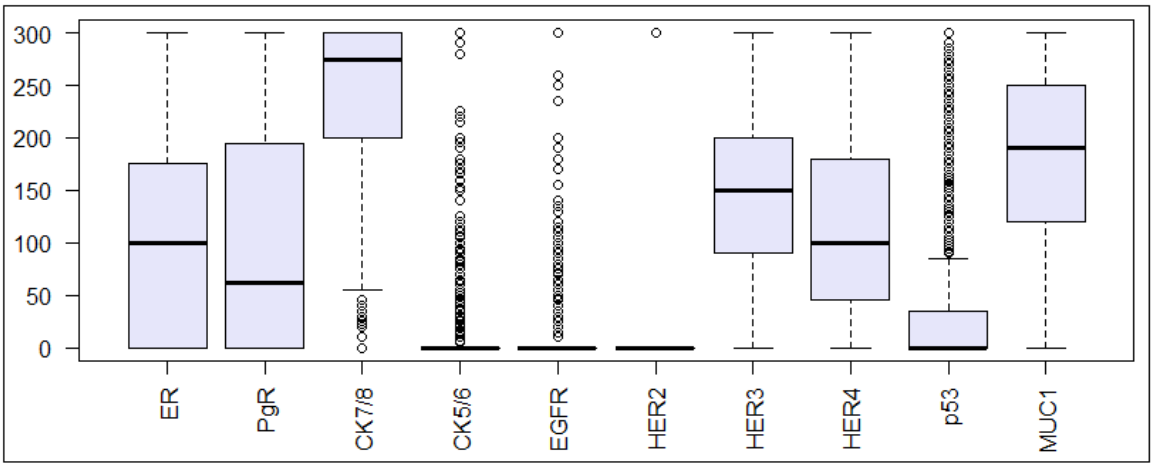


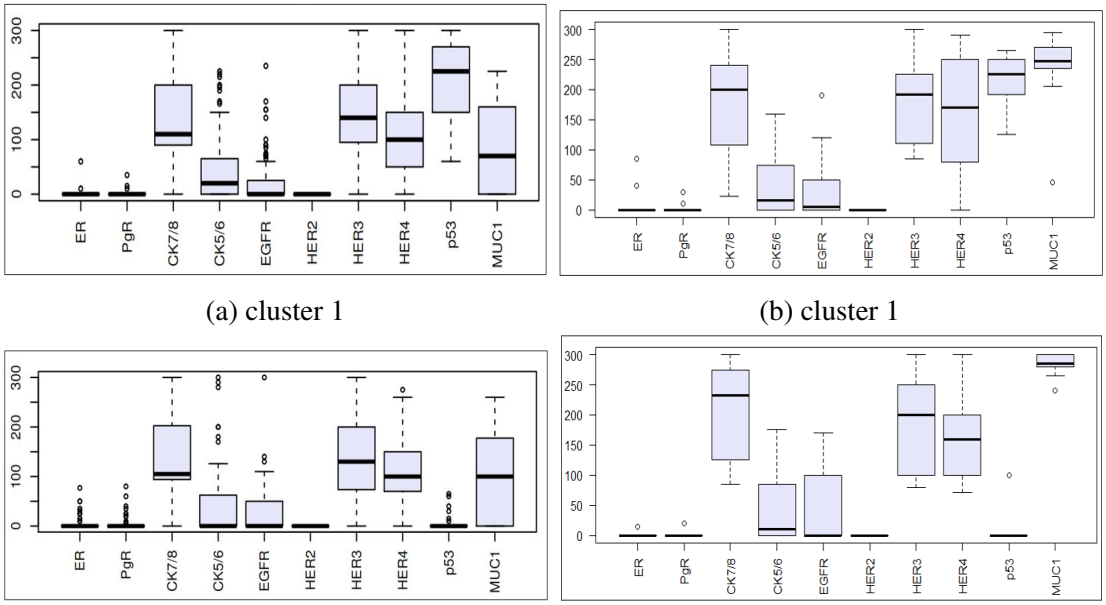
Figure 3.3 Boxplot of 1073 patients from the Nottingham Breast Cancer Series.

Among 18 patients, six can be classified in class 4, 11 in class 5 and the one remaining in class 1. Figure 3.3 shows the boxplots of the original 1,073 breast cancer patients where the y-axis represents the H-score and the x-axis represents the ten protein markers.

Figure 3.4 compares the boxplots of the original classes 4 and 5 with those from the current work. It can be seen that previous and new methodologies generate similar boxplots, validating the new results. The 'MUC1' biomarker has a different mean and a different range of values in the new classes 4 and 5. The mean is high in both the new classes. Similar behaviour is observed for the protein 'CK7/8' i.e. the mean is high and the range is different in both the new classes 4 and 5. The difference in the values of both the biomarkers could be the reason for misclassification of these patients in the original work by Soria et al. [10]. The strong affinity for class 4 and class 5 is validated by the range and the mean of the remaining eight biomarkers.

3.2.6 Summary

To tackle the breast cancer heterogeneity, the identification of groups is extremely important to tailor and choose the ideal treatment. In this section we have presented a potential solution for classification of difficult to classify (i.e. 'not classified') patients. The work improves the classification accuracy with in the Nottingham Breast Cancer dataset by classifying 18 more patients into one of the seven core breast cancer classes. The solution forms the basis for an improved classification framework for unlabelled datasets, presented in the next section.



(a) cluster 1 (b) cluster 1

(c) cluster 2 (d) cluster 2

Figure 3.4 Boxplots of patients (a) Class4 of original classification (b) Class4 of new methodology (c) Class5 of original classification (d) Class5 of new methodology.

3.3 Two-stage Data Classification Framework

While combining the clustering algorithm results using ensemble methods, a number of data are left unclustered, explained in the previous section using a case study on Nottingham Breast Cancer dataset. Researchers have focused on making ensemble clustering combination as good as possible, so to leave the smallest number of data unclustered [8, 72]. In this section, we build on this work and present a step-by-step framework for the identification of groups and efficient data distribution within them. Unlike other approaches in literature [8] which focus on improving the ensemble clustering, the framework proposed in this thesis employs two stages: ensemble clustering followed by ensemble classification. The first stage is ensemble clustering in which the core groups are identified and subsequently a subset of data (for which the cluster can be identified) is assigned to one of the clusters. It can be observed that the first step is similar to other methods proposed in the literature and similar to them a number of data points are left unclustered. In order to have the maximum amount of data clustered in one of the identified groups, a second stage is added. The idea is to now see this as a classification problem, i.e. using the clustered data as the training set to train the classifiers and use the unclustered data as the test set. The introduction of the ensemble classification step results in a novel classification framework.

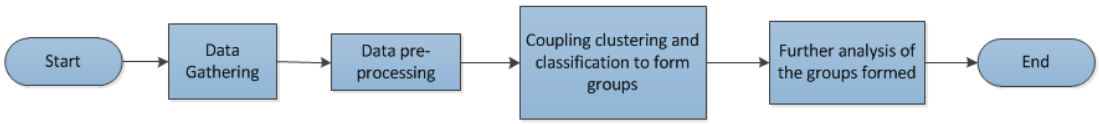


Figure 3.5 Abstract steps in the proposed framework

The two stages can be broken into multiple steps. An abstract view for the steps of the framework is shown in Figure 3.5. The first step in the framework is data gathering, followed by cleaning in the pre-processing unit. The processed data are passed to the 'Coupling ensembles of Clustering and classification' unit which is itself broken down in further sub-steps, each discussed in the next section. The final clustering in groups is inspected in the further analysis step. The units of the framework are explained in the following subsections.

3.3.1 Data Gathering

The first step in the framework is aggregation of the data. Data can come in different shapes and formats, but the process and protocols to collect the data are out of scope of this study.

3.3.2 Data Pre-processing

The next step is data pre-processing, i.e. data cleaning, which enhances the power of algorithms to identify the patterns inside the datasets. In order to fully understand the data under inspection descriptive statistics such as mean, median, range, standard deviation, etc. can be computed. This might also help to locate inconsistencies among the data. Tuples which have missing values can be deleted, and the tuples attributes could be made homogeneous, i.e. all numeric or text. Other ways for processing could be normalisation of values using methods such as min-max, z-score, etc. [142]. These steps produce the final set of data points to be used.

3.3.3 Coupling ensembles of Clustering and Classification

Coupling ensembles of Clustering and Classification is composed of multiple sub-steps, as depicted by a detailed flowchart in Figure 3.6. In this unit, the first sub-step is Ensemble Clustering Unit, explained in the following subsection.

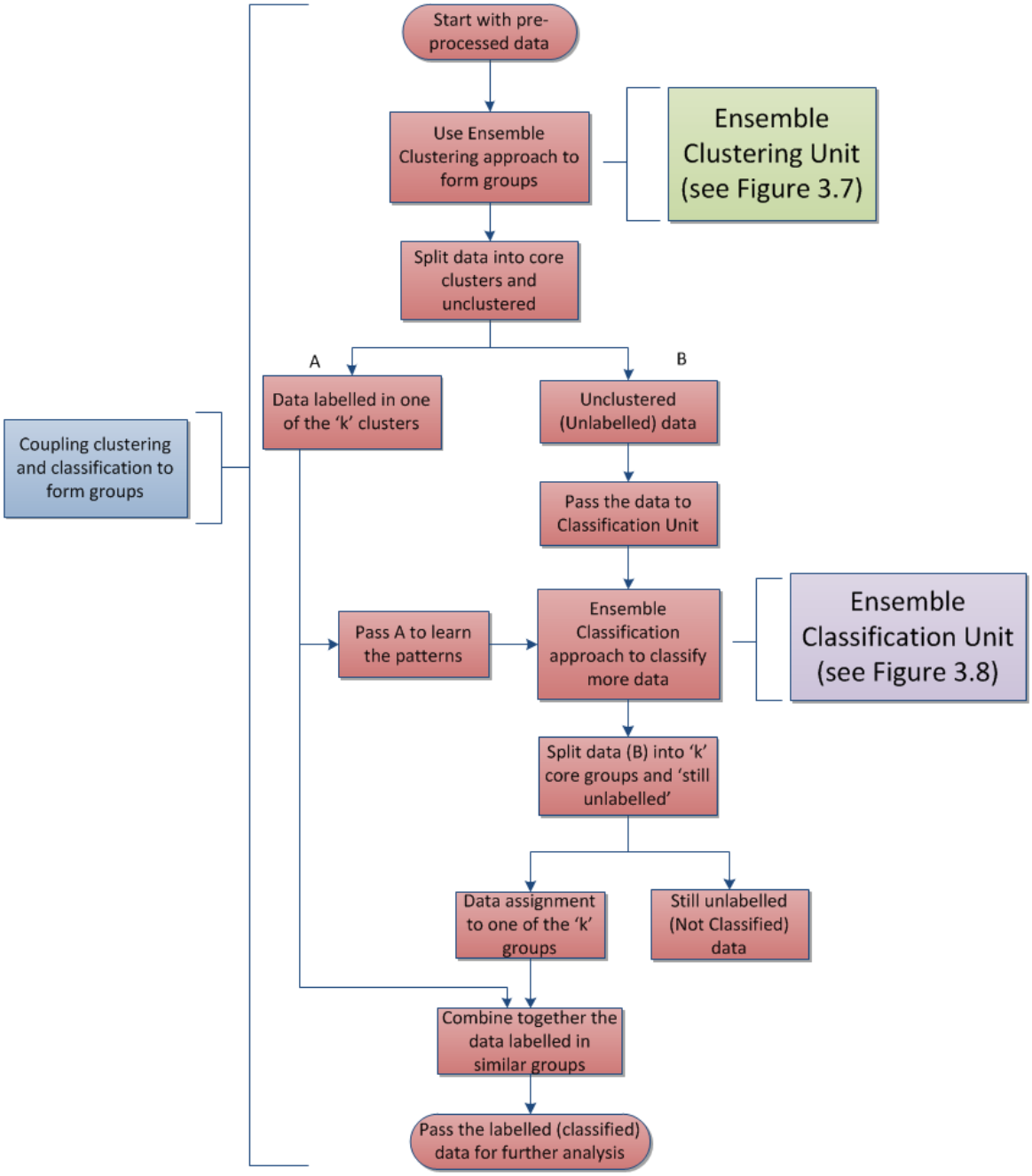


Figure 3.6 Flowchart of the sub-steps in 'Coupling Clustering and Classification to form groups' step from the abstract view. 'Ensemble clustering (Clusterfusion) to form groups' and 'Ensemble Classification approach to classify more data' are further subdivided into smaller steps explained in Figures 3 and 4 respectively.

Ensemble Clustering (Clusterfusion) Unit

This section elucidates the construction of groups and initial data distribution from the dataset by consensus among the results of the two clustering algorithms. The cluster ensemble methodology looks similar in principle to equivalent approaches in classification, namely ensemble methods like bagging and boosting, but it presents additional problems [72]. Firstly, matching clusters between algorithms is not straightforward as different clustering algorithms may generate different numbers of groups and moreover the optimal number of groups may be unknown. Another common problem is that of unclustered data left while combining the results of the clustering algorithms.

To determine the optimal value of number of clusters 'k', we suggest to run cluster validity indices. Set the minimum and maximum values for k, then at each iteration run the clustering algorithms and the validity indices, changing the input k between the two set values. If the indices suggest different numbers of clusters, then the minimum sum of ranks can be taken as the optimal value of k [8]. This information about k is now passed back to cluster the data using the same two clustering algorithms. To verify the ensemble solution, Swift et al. [143] suggested the use of the kappa index. The agreement between the equivalent clusters returned by the different clustering algorithms is evaluated using the Cohen's unweighted and weighted kappa index [144]. The index statistically evaluates the agreement between the two clustering algorithms which clusters n data among k groups, and return a value between 0-1 with zero being absolute disagreement and one being complete agreement. Thus, this process returns the most stable cluster size.

The groups obtained by the different clustering algorithms are merged with the goal of assigning the same set of objects to identical clusters, a process known as 'correspondence problem'. The correspondence among the groups is achieved by aligning the group labels to get maximum agreement and therefore form core groups composed of the data with the same group label. These steps of the ensemble clustering unit are represented by a flowchart in Figure 3.7. There might be some data that present mixed classification while dividing the data points into k common groups. These remaining data are therefore called unclustered (unlabelled). To tackle the unclustered data we introduce the use of Ensemble Classification on top of clusterfusion and forward the unclustered data to the 'Ensemble Classification Unit'.

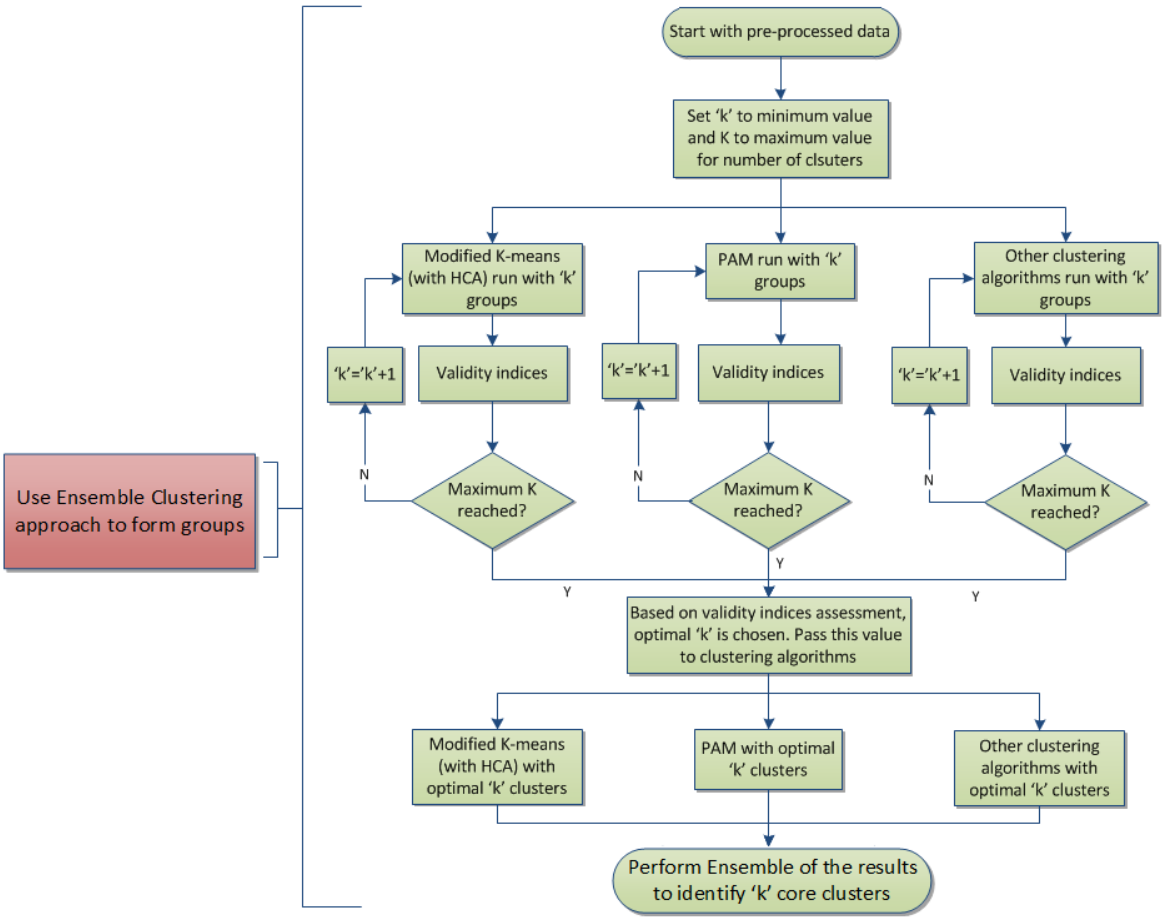


Figure 3.7 Flowchart for Consensus Clustering Unit shows the pictorial representation of the order of steps to form groups in a dataset and subsequently divide the data in these constituted groups.

Ensemble Classification Unit

The ensemble classification unit is proposed to refine the clustering results by targeting the data that present mixed classification. The reasons behind introducing the ensemble classification in the framework are to improve solutions, to reuse knowledge and to select novel models by changing classifiers. Both the classification algorithms first learn the classification rules from the previously clustered data (training set), and then use the unclustered one as the test set with the goal to assign them to one of the previously identified 'k' groups. Decision-level fusion using unanimous voting is employed to merge the classification results. The reason for using this method is the choice of real world application to assign breast cancer patients to biological groups, which demands high confidence in the results. As Unanimous voting

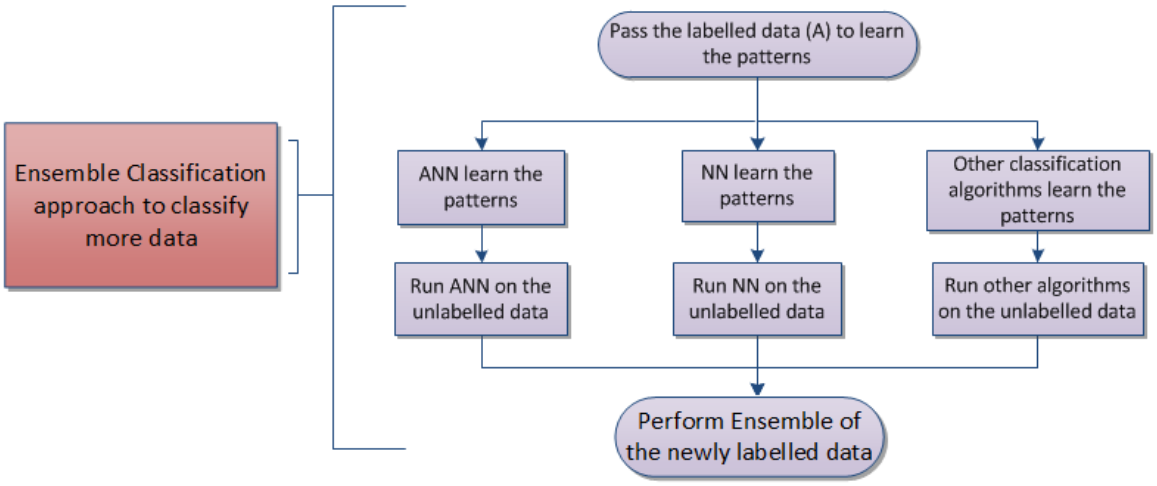


Figure 3.8 Flowchart for the order of steps in Consensus Classification Unit to classify more data in groups constituted in the clusterfusion step.

considers the consensus across all the algorithms in the ensemble, it generates the highest confidence in results among all the ensemble methods.

Two principles were considered while performing the consensus: i) include as many classifiers in the ensemble as possible, and ii) assign as many data to the core groups as possible. Strict application of the principles may result in few data assignments, leading to conflicts between them. The possible solution for this problem can be: i) to use other methods of ensemble classifiers such as majority voting, weighted voting, stacked generalisation, etc. ii) to use a trade off between the two principles. However, this problem is currently overlooked as high output confidence is desired in this work.

After ensemble classification, more data are assigned to one of the groups, and the remaining data are still unlabelled or 'Not classified'. The data clustered after the clusterfusion step are now combined with the newly clustered data in the same groups to achieve the final clusters. The newly clustered data needs further analysis and verification, which is discussed in the following section. The detailed flowchart for the order of steps in this unit is shown in Figure 3.8.

3.3.4 Further Analysis of the Groups Formed

To evaluate and validate the characteristics of the data assigned after ensemble classification, visual tests, statistical tests and cluster quality assessment need to be performed. The distribution of each attribute in a group after clusterfusion was compared to the one after ensemble classification using boxplots (a method of visually displaying the distribution of the

data using median, quartiles and outliers [145]) and the Davies-Bound index (internal cluster quality assessment index). The statistical test used for the comparison is Mann-Whitney-Wilcoxon, a non-parametric version of the t-test, to check whether the two distributions come from the same population. The level of significance is usually chosen at 0.05 [146].

3.4 Summary

The identification of characteristic groups and subsequent classification of data points among them is of prime importance as part of understanding and exploring unlabelled datasets. A specific issue within these datasets is the classification of difficult to classify data points. For example, classification of patients across biological groups within the breast cancer datasets. Recent studies have used an ensemble of multiple clustering algorithms to elucidate the most characteristic groups from these unlabelled datasets. However, the combination of various clustering methods resulted in a number of data points as unclassified. Thus, in this chapter a novel two stage classification framework is introduced by adding an ensemble classification stage after an ensemble clustering stage.

To motivate the need for an improved classification framework, a case study on Nottingham Breast Cancer dataset was presented. Initially 38 out of 1,073 were 'not classified' using the previous frameworks from literature. Using the novel two-staged classification framework an improvement in classification was observed, as only 18 remained 'not classified'.

To demonstrate the performance of this new classification framework, it needs to be empirically evaluated. Thus, in the next chapter three case studies (including Edinburgh breast cancer dataset, several benchmark datasets from UCI repository and Microlise's heavy goods vehicle dataset) are employed to systematically evaluate the two-stage classification framework. Moreover, the application on these case studies will also demonstrate the generality and robustness of the classification framework.

Chapter 4

Validating Robustness of the Classification Framework

This chapter validates the practical impact and robustness of the two stage classification framework by empirically evaluating it across multiple real-world applicaitons. Therefore, this chapter will also address objective 4 stated in Section 1.2, i.e. real-world application of the new classification framework. It is important to note that in the last chapter we used a real world case study to demonstrate the *need* for a new classification framework, where as, this chapter will present the three case studies to *evaluate* the two stage classification framework. The structure of the chapter is as follows: Section 4.1 presents the first case study on Edinburgh breast cancer datasets for biological group stratification. The second case study is to examine multiple standard datasets from the UCI repository for stratification of their characteristic groups, demonstrated in Section 4.2. Lastly, Section 4.3 will present the third and final study on Microlise's heavy good vehicle dataset for driving stereotype identification.

4.1 Case Study 1: Biological Group Stratification in Edinburgh Breast Cancer Dataset

4.1.1 Problem Overview

In the domain of breast cancer, the identification of groups is extremely important for clinical experts to determine the best treatment outcome. Previously, ensemble clustering algorithms have been used to elucidate core biological groups in a breast cancer dataset. However, the

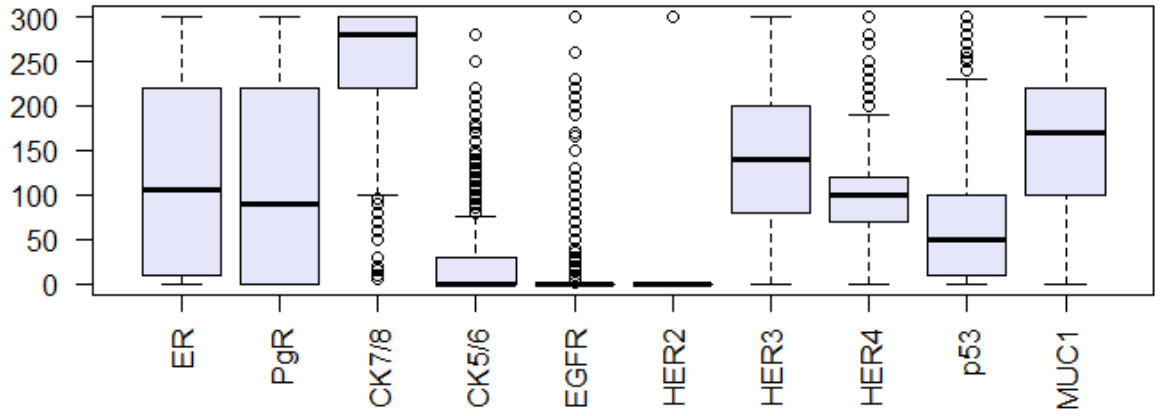


Figure 4.1 Boxplots of 478 patients in the Edinburgh Dataset

methodology resulted in a number of unclustered patients, i.e. low classification. Thus, in the previous chapter we introduced the use of ensemble classification and presented a novel framework for improved data distribution in the identified groups. As stated above, in this section we present the first case study on Edinburgh breast cancer dataset for improved identification of core groups, data distribution in these groups and improvement in final classification by re-classifying the unclustered data.

4.1.2 Dataset Description

The real world data used was a breast cancer dataset provided by the Edinburgh Research Centre. The Edinburgh Breast Cancer Series consists of 885 patients treated with breast conservation surgery, axillary node sampling or clearance, and whole breast radiotherapy [104]. Each of the patient's tumour samples were tested against ten biomarkers using immunohistochemistry, which resulted in a score for each biomarker ranging from 0 to 300. The biomarkers were: Estrogen Receptor (ER), Progesterone Receptor (PgR), Cytokeratin 7/8 (CK7/8), Cytokeratin 5/6 (CK5/6), EGFR, c-erbB2 (HER2), c-erbB3 (HER3), c-erbB4 (HER4), p53 and Mucin1 (MUC1) [10], [106].

The next step after selecting the dataset is data processing. In the Edinburgh breast cancer dataset, the biomarkers ER, PgR and HER2 had no missing values for all the patients while other biomarkers had some missing entries. All the tuples with at least one missing value were deleted, reducing the number of patients down to 478. Since all the biomarkers were scored on a 0 to 300 range, normalisation was not required. The cleaning of the cancer dataset was followed by the computation of statistical measures (namely median, range and outliers)

Validity Index	PAM	K-means
Marriot	4	4
Calinski	4	4
Scott	4	4
TraceW	4	3

Table 4.1 Number of clusters suggested by different validity indices

and boxplots. Figure 4.1 shows the boxplots of the 478 breast cancer patients where the y-axis represents the values for each biomarker and the x-axis represents the ten biomarkers.

4.1.3 Results

After the descriptive analysis, the next step is to choose the number of groups by employing PAM and K-means clustering algorithms for k = 2 to 15. The cluster results obtained were passed to the cluster validity indices to assess the quality of the clustering output and hence determine the optimal number of groups. Each validity index (Marriot, Calinski, Scott and TraceW) was run separately on each clustering algorithm, with an associated rule to choose the best number of clusters, already shown in Table 2.1. All the Cluster Validity indices suggested four to be the optimal number except TraceW when run after K-means, for which the optimal number was three. The results of the validity indices computation for both the algorithms are shown in Table 4.1. The minimum sum of ranks was taken and four was chosen as the optimal number.

Both the clustering algorithms were again run on the dataset for four groups. The correspondence between results of the two clustering algorithms were statistically compared using the unweighted and the weighted Cohen's kappa index. The weights of the weighted kappa index were computed based on the degree of disagreement between the groups, i.e. the higher the weight, higher the disagreement. The unweighted kappa index value was 0.85 and the weighted kappa index value was 0.93. These two values indicate a very high agreement between the two clustering algorithms for the four groups formed. The cluster labels of both the clustering algorithms were aligned in order to have the same data assigned to the cluster with the same label from the two methods. The distribution of patients among the four classes is shown in the first two rows of Table 4.2. Clusters 1, 2, 3 and 4 represent the four identified groups. Ensemble clustering grouped 426 patients among the four common clusters and left 52 unclustered (row three of Table 4.2).

The next step was to train the two classifiers on the 426 clustered patients, run them on the 52 unclustered patients as test set and perform the class level fusion of results. The first

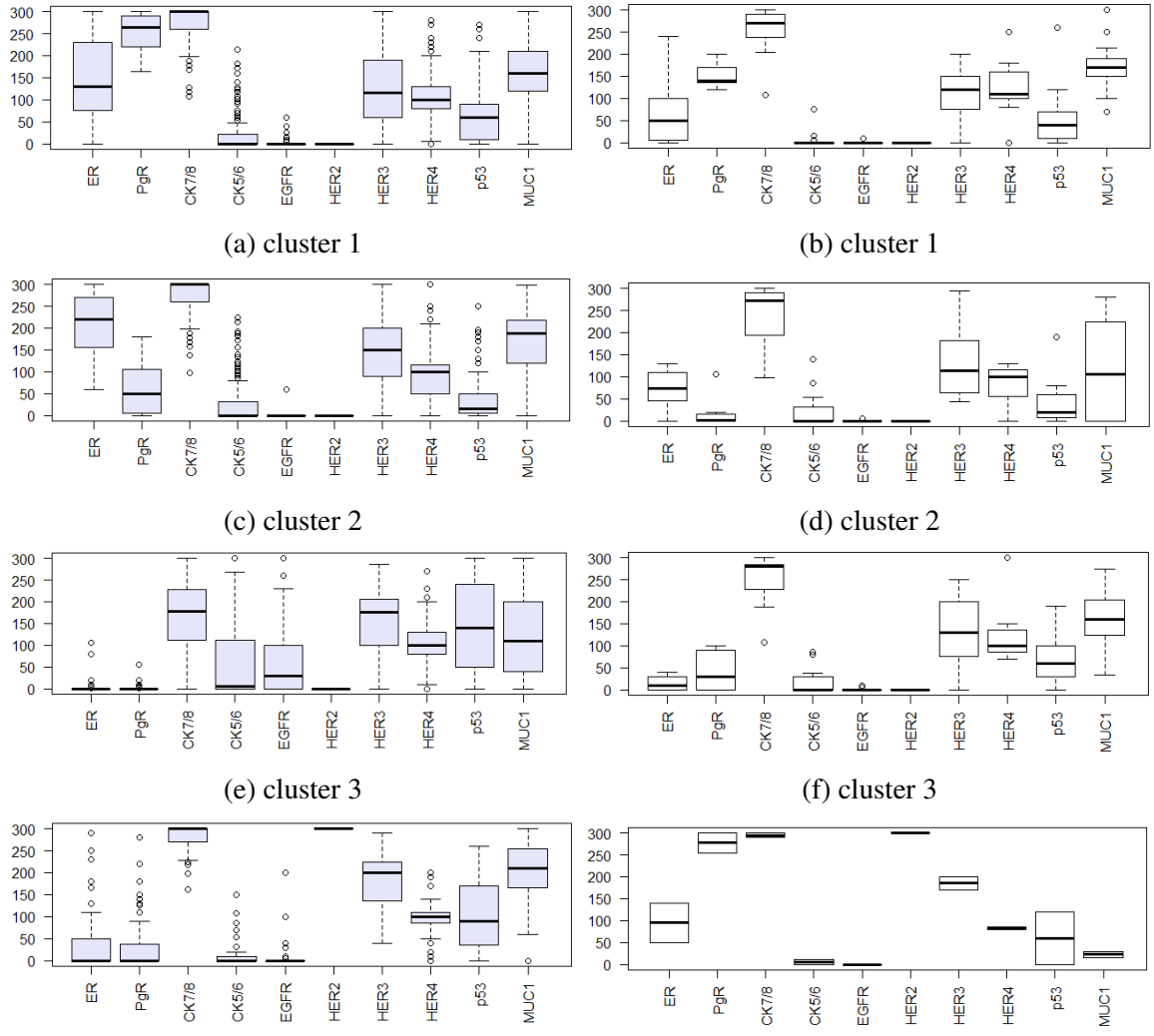
Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Unclustered
K-means	168	141	120	49	-
PAM	160	190	81	47	-
Ensemble Clustering	157	141	81	47	52
Ensemble Classification (additionally assigned)	13	12	11	2	14
Coupling ensemble clustering with ensemble classification	170 (35.57%)	153 (32%)	92 (19.25%)	49 (10.25%)	14 (2.93%)

Table 4.2 Division of patients in clusters

classifier used was five layered ANN network (having three hidden layers), with a sigmoid activation function to model the weighted sum. The ANN was tested with multiple models, but it achieved best results for network having 7-10-7 nodes for the hidden layers, with 200 epochs. The second classifier was the Nearest Neighbour with an Euclidean distance function. For both ANN and Nearest Neighbour, we used a 10 fold cross validation approach. An additional 38 patients were assigned to the common groups whose distribution is shown in row four of Table 4.2. Combining ensemble clustering and ensemble classification, a total of 464 patients were clustered in one of the groups and the final distribution among the clusters is shown in the last row of Table 4.2.

The data assigned to groups after ensemble classification were visually verified using boxplots. Figure 4.2 compares the boxplots of the biomarkers for patients clustered in four groups after ensemble clustering with the boxplots for the patients subsequently assigned to the groups among the unclustered ones. The boxplots on the left side, i.e. (a),(c),(e) and (g), are of 426 patients in clusters 1, 2, 3 and 4 respectively after ensemble clustering. Similarly, boxplots on the right side, i.e. (b),(d),(f) and (h), are of 38 more patients assigned after ensemble classification in clusters 1, 2, 3 and 4 respectively. Of these 38, 13 were assigned to cluster 1, 12 to cluster 2, 11 to cluster 3 and two to cluster 4.

It can be observed that the boxplots of the patients after ensemble clustering are very similar to the ones after ensemble classification for corresponding groups implying differences between the medians may not be significant. Few biomarkers differ, as in cluster 1 after ensemble clustering, the medians of ER and PgR are in the range of 100 to 150 and 250 to 300 respectively, while after ensemble classification the median is around 50 for ER and 150 for PgR. Cluster 2 shows the variation only for one biomarker, ER, with its median in the range of 200 to 250 for patients after ensemble clustering and being in the range 50 to 100 after ensemble classification. Cluster 3 also shows a difference for only



(a) cluster 1

(b) cluster 1

(c) cluster 2

(d) cluster 2

(e) cluster 3

(f) cluster 3

(g) cluster 4

(h) cluster 4

Figure 4.2 Boxplots of biomarkers. (a), (c), (e), (g): for patients in clusters derived after ensemble clustering. (b), (d), (f), (h): for additional patients classified after ensemble classification.

Biomarker	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ER	0.004	$2.34\text{x}10^{-7}$	0.125	$8.02\text{x}10^{-6}$
PgR	$5.69\text{x}10^{-9}$	0.003	0.001	$6.89\text{x}10^{-5}$
CK7/8	0.061	0.072	0.005	0.681
CK5/6	0.158	0.539	0.214	0.84
EGFR	0.937	0.988	0.0004	0.467
HER2*	1	1	1	1
HER3	0.587	0.074	0.356	0.879
HER4	0.25	0.962	0.751	0.202
p53	0.438	0.766	0.0325	0.39
MUC1	0.637	0.055	0.082	0.024

*The reason for the p-value for HER2 to be equal to 1 is because the HER2 H-score for all patients in the data was either 0 or 300, and it did not follow exactly a continuous range of values as for the other biomarkers.

Table 4.3 The p-value comparison for patients after ensemble clustering with ensemble classification

one biomarker, CK7/8, with the median being in the range of 150 to 200 for patients after ensemble clustering and in the range of 250 to 300 for patients after ensemble classification. The differences among biomarkers in cluster 4 are among ER, PgR and MUC1: the medians after ensemble clustering are almost 0 for ER and PgR and just more than 200 for MUC1, while for patients clustered after ensemble classification the value of the median for ER is almost 100, between 250-300 for PgR, and that of MUC1 is below 50. For other biomarkers in all the classes, the medians are almost equal. The reported differences in the values of the biomarkers could be the reason that patients grouped after ensemble classification were originally unclustered after ensemble clustering.

We now check whether the aforementioned differences are statistically significant, for which we compare the distributions of markers in the newly formed groups with the groups from ensemble clustering, using the Mann-Whitney-Wilcoxon test. Table 4.3 shows the p-values comparison of 426 patients after ensemble clustering with 38 after ensemble classification for all the clusters and markers. The level of significance was set at 0.05. In both clusters 1 and 2, it can be seen that ER and PgR have statistically different distributions. In cluster 3, significant differences are observed for PgR, CK7/8, EGFR and p53. In cluster 4, the biomarkers ER, PgR and MUC1 are significantly different, similar to the boxplots observation. Based on the boxplots and statistical results, the new patients added after ensemble classification could be assigned to one of the groups.

To understand the significance of adding the ensemble classification layer on top of the ensemble clustering layer, the Davies Bound (DB) [147] internal cluster validity index is

Assigned Cluster	13 patients from Cluster 1	12 patients from Cluster 2	11 patients from Cluster 3	2 patients from Cluster 4
Cluster 1	**1.3840**	1.4004	1.3933	**1.3736**
Cluster 2	1.4339	**1.3877**	1.4297	1.3835
Cluster 3	1.4138	1.3991	**1.3915**	1.3893
Cluster 4	1.4016	1.4442	1.3941	**1.3749**

Table 4.4 DB index value of each subgroup of patients assigned after ensemble classification, compared by assigning to all the other remaining clusters.

used, which is one of the most stable and commonly used indexes in the literature. The DB index computes the ratio of intra-cluster and inter-cluster distance of the points (i.e. patients), therefore, the lower this value the better the clustering. Since additional patients are added to the clusters, the value of DB will slightly increase. As the value of the DB index increases, the quality of the final classification decreases and thus the lower the increase, the better the classification. Therefore, to study the quality of clusters, the DB index values of each subgroup of patients assigned after ensemble classification were compared to the values of the DB index obtained if they had been assigned to each of the other clusters.

The DB index value of the overall classification after ensemble clustering is 1.3732. Table 4.4 shows the DB index value of each subgroup of patients assigned to each cluster. Each column represents the cluster predicted for a set of patients after ensemble classification, i.e. third row of Table 4.2. Each row contains the DB index of the assigned cluster. It can be observed from that for Clusters 1, 2, and 3, ensemble classification assigns patients to the best possible group, while for cluster 4, the two additional patients were assigned to the second best group. Overall, the DB index supports the distribution of patients after ensemble classification. It was also observed that at least one of the clustering algorithms did assign the elements to the same class as obtained after ensemble classification. Therefore, the use of the DB index supports this, and the clusters in which the patients are added after ensemble classification is the best possible.

4.1.4 Clinical Assessment

We presented the results to clinical experts for their interpretation. Based on a review of the results by clinical expert, we can be confident in assigning the patients after ensemble classification to the biological groups identified after ensemble clustering stage. The final clusters (after combining ensemble clustering and ensemble classification) represent the four biological groups reported in Figure 4.3, shown using a tree format. According to available

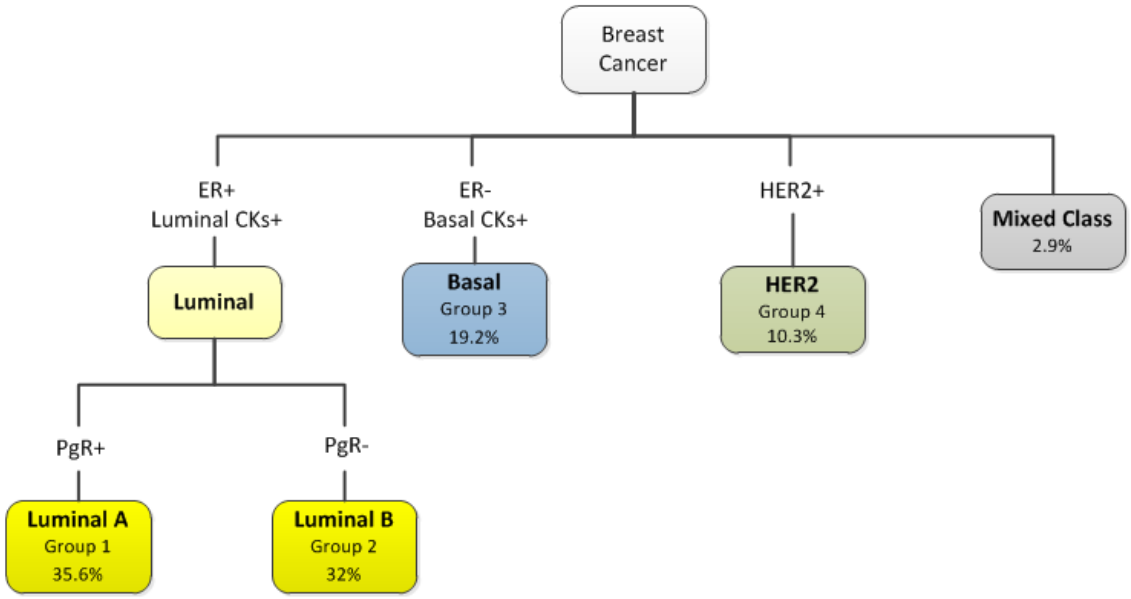


Figure 4.3 Breast Cancer Groups with representative characteristics

pathological information ER and PgR levels define Luminal groups. The two groups are characterised by positive levels for ER values but different PgR levels, i.e. positive and negative respectively. Thus, Clusters 1 and 2 represent Luminal A and Luminal B biological groups respectively. Regarding those patients assigned to groups after ensemble classification, clinicians stated that, although values may differ in some biomarkers, the patterns of the defining biomarkers need to be consistent. In Clusters 1 and 2 patients after ensemble classification follow the same trend, i.e. ER positive/PgR positive and ER positive/PgR negative respectively.

Cluster 3 represents Basal group with p53 altered, characterised by high levels of p53 and low levels of ER. Patients in cluster 3 after ensemble classification have high levels of p53 and low levels of ER similar to the patients after ensemble clustering. Cluster 4 characterise HER2+ biological group, having high values of HER2 biomarkers. The patients after ensemble clustering have low levels of biomarker ER, called HER2/ER- group. However, the patients assigned to cluster 4 after ensemble classification have high levels of ER (called HER2/ER+), which could be the reason for them being unclustered. Since the number of patients after ensemble classification are very small i.e. 2, thus they are labelled under one biological group called HER2+.

4.1.5 Discussion

While exploring the literature for the division of patients in biological groups using ensemble clustering, the need for improving the clustering results was identified. For example, Soria et al. [8] used ensemble clustering on a breast cancer dataset but could only classify 62% of the patients in biological groups. Kellam et al. [114] also used ensemble clustering (on a different domain though) and could only classify 55% of the patients. Thus, in this thesis we use an improved two stage classification framework targeting the difficult to classify data (in an unlabelled dataset). K-means (with HCA) and PAM clustering algorithms have been used in first stage of the framework as they are two of the most commonly used clustering algorithms with good performance [148]. Subsequently, an ensemble classification was used to distribute more patients, which were previously unclustered. Two classifiers (ANN and the Nearest Neighbour) have been utilised for the ensemble classification phase [13], as together they have consistently resulted in good performance throughout the literature [149] and cover advantageous properties required for good ensemble classification such as non-linearity, non-parametric, memory based learning, etc. The application of the framework on the Edinburgh Series identified four possible biological classes and distributed the patients among them. From Table 4.2 it can be observed that after employing ensemble clustering 89% of the patients were assigned to one of the groups, and the classification jumped to 97% after adding ensemble classification stage.

Additionally, Soria and colleagues in two consequent works ([8] and [10]) proved the existence of three major biological groups (actually seven in total, three being the main groups and four of them being sub-groups): Luminal, Basal and HER2. Therefore, by comparing the classification results of Edinburgh dataset with the available pathological information (and based on the characteristics), we have obtained similar biological groups as in the literature. The patients were distributed among four biological groups using the framework, as cluster 1 can be labelled as Luminal A, cluster 2 as Luminal B, cluster 3 as Basal, cluster 4 as HER2+ and the remaining patients as Mixed Class (Figure 4.3).

4.1.6 Summary

In the domain of breast cancer, the identification of groups is extremely important for clinical experts to determine the best treatment outcome. In this section we employed the novel two stage classification framework (introduced in the last chapter) on the Edinburgh breast cancer dataset (for improved data distribution in the identified groups). First the ensemble of multiple clustering algorithms was been used for the extraction of characteristic biological

groups from a breast cancer dataset and initial distribution of patients in them. 426 data out of 478 were clustered among one of the four groups after ensemble clustering, verified by high kappa index values. Next, the ensemble classification algorithm used the 426 patients in the four groups to learn the patterns. After this step, an additional 38 patients out of 52 were added to one the four groups. The final data distribution in groups was achieved by combining the patients distributed after ensemble clustering with the patients distributed after ensemble classification. Thus, a total of 464 patients were clustered using a two stage methodology.

We verify the data grouped after ensemble classification by comparing it with the data clustered after ensemble clustering, using statistical (Mann-Whitney-Wilcoxon) [39] and visualisation (boxplots) tests. We also tested the cluster quality using the Davies-Bound index, to verify that patients were added to the most appropriate group [147]. The results show from Table 4.2 show that with the addition of ensemble classification, the framework improves the ensemble clustering results.

An additional case study has been done to analyse the importance of biomarker Ki67 in identifying Breast Cancer treatment groups. This study was performed for a new dataset collected using the Reverse Phase Protein Micro-array (RPPA) approach. The study on both the RPPA and the Edinburgh dataset are performed to identify the biological treatment groups with a breast cancer dataset. The study in not included this chapter as it was performed differently to all the three case studies presented in this chapter. Moreover, the material is too short to be presented as a standalone chapter. We have therefore included the results in Appendix A. Next section presents the case study on multiple datasets from UCI repository.

4.2 Case Study 2: Characteristic Group Stratification in Benchmark Datasets

4.2.1 Problem Overview

In the second case study, the framework is employed to multiple standard datasets from the UCI machine learning repository [150]. This case study is performed to show the generality and robustness of the framework. The classification results in the previous section were verified using multiple tests including statistical, visual and clinical. However, the datasets used in this section already had ground truth available. Upon classification the results were verified against the ground truth, uncovering interesting results, discussed below.

4.2.2 Dataset Description

Five UCI benchmark datasets [150] including two standard medical datasets were used to evaluate the framework i.e. external validation. These five datasets were selected as they contain different number of classes with a broad range of features, shown in Table 4.5. The framework was run on all the datasets by removing the ground truth from them. For each dataset, after distributing the data among the groups using the classification framework, the results were compared with the ground truth.

4.2.3 Results and Discussion

Table 4.6 shows the results of the framework on the UCI datasets. The first row contains the number of data clustered in the groups after ensemble clustering represented by 'Number of data clustered after Ensemble Clustering'. The second row contains the number of data correctly grouped after ensemble clustering by comparing with the ground truth represented by 'Number of data correctly clustered after Ensemble Clustering'. The third row contains the number of data clustered in the groups after ensemble classification and the fourth row the number of data correctly grouped after ensemble classification represented by 'Number of data clustered after Ensemble Classification' and 'Number of data correctly clustered after Ensemble Classification', respectively. 'Total data correctly clustered' contains the sum of number of data correctly clustered after ensemble clustering and ensemble classification. The last two rows contain the percentage of data correctly clustered after ensemble clustering and the percentage of total data accurately clustered respectively. The increase in percentage shows the improvement in results upon adding ensemble classification stage, as discussed in the next section.

To measure the effectiveness, robustness and generalisation of the proposed framework, in second case study it was run on standard datasets by removing their ground truth. From Table 4.6 comparing the percentage of data correctly clustered after ensemble clustering

	Iris	Wine	Ecoli	Dermatology	WBC*
Number of Classes	3	3	8	6	2
Number of Features	4	13	7	33	9
Total Number of Data	150	178	336	366	683

*Wisconsin Breast Cancer

Table 4.5 Descriptive features such as number of classes, number of features and total data in the benchmark datasets

Description	Iris	Wine	Ecoli	Dermatology	WBC*
Number of data clustered after Ensemble Clustering	134	170	334	351	681
Number of data correctly clustered after Ensemble Clustering	133	161	289	335	655
Number of data clustered after Ensemble Classification	16	8	2	15	2
Number of data correctly clustered after Ensemble Classification	11	7	2	15	2
Total Data correctly clustered after combining Ensemble Clustering and Ensemble Classification	144	168	291	350	657
Percentage of Data correctly clustered after Ensemble Clustering	88.67%	90.5%	86%	91.5%	95.9%
Percentage of Total Data correctly clustered after combining Ensemble Clustering and Ensemble Classification	96%	94.4%	86.6%	95.6%	96.2%

*Wisconsin Breast Cancer

Table 4.6 A comparison of distribution of data among the groups for the benchmark datasets.

(seventh row) and data correctly clustered after the novel two stage method (eighth row), it can be observed that there is a significant improvement in the classification results by adding an additional ensemble classification stage. This improvement is consistent across all the datasets. Thus, the experimental results show that the proposed framework effectively finds groups in the datasets and distributes the data points in them (results validated by comparing with the ground truth). Although the framework does not achieve the highest accuracy on standard datasets in comparison to other supervised classification algorithm accuracies in literature, these differences are small and acceptable. Moreover, these differences occur as the framework takes the unlabelled data as input while the supervised algorithms learn the classification rules on the ground truth in the training set. Thus, a trade-off can be considered to identify groups in the datasets with no class labels and to get most of the data correctly clustered.

4.2.4 Summary

This case study expands the initial idea of employing the ensemble classification method in a breast cancer identification framework, and its focus is not on the number of clusters formed but on the overall methodology to refine clustering results. Subsequently, in this case

study we presents the application of the framework on several standard datasets from the UCI repository. The biggest advantage of the proposed framework lies in its adaptability, which enables customisation of each component (algorithms can be changed depending upon the dataset) according to need and at each stage i.e. a *robust* approach. Although in this section we have used specific clustering algorithms, classification algorithms, ensemble methods and cluster validity indices, all of them could be replaced by other similar algorithms.

Upon distributing the data for the UCI datasets, the final classification of the data in the groups was verified with already present ground truth for them. Although the final accuracy was equivalent to what is achieved by other classifiers in the literature on the supervised algorithms, these difference are acceptable. The results on the standard dataset show the generality, robustness and effectiveness of the framework, particularly for those datasets without labels.

To further evaluate the classification framework, in the next section another practical case study is used, which is non-medical and real world. This case study is used to demonstrate that the framework can be used across different applicaitons.

4.3 Case Study 3: Identifying Heavy Goods Vehicle Driving Styles in the United Kingdom

4.3.1 Problem Overview

In the third case study, we demonstrate the practical impact of the classification framework with application on another real-world dataset. In order to do so, we focus on identifying the driving stereotypes amongst the UK Heavy Goods Vehicle (HGV) driving community by analysing a telematics dataset of driving incidents. 21,193 drivers were analysed in this case study, collected in real-time. The next section discusses the background and motivation for the need of using the classification framework in this application.

4.3.2 Background and Motivation

A large proportion of private and public sectors in the United Kingdom (UK) rely on heavy goods vehicles (HGVs) transport for procurement and delivery of goods and services. Statistics for the UK between October 2014 and September 2015 show that HGVs delivered 1.63 billion tonnes of freight within the UK, and 8.5 million tonnes were imported and exported [151]. Due to the importance of HGVs in the country's economy, there are great

efforts being employed to reduce their incident numbers [152, 153]. Such incidents, which incur in significant losses, are caused by the characteristics of the vehicle, road, weather conditions, defiance of traffic rules, company policies and mostly by driving behaviour. In order to mitigate these mishaps and their consequences, companies and researchers have invested time and resources to improve the technology to manage HGV fleets. Fleet management includes a range of activities, such as vehicle maintenance, telematics, driver monitoring, health and safety inspections, etc. Most of these practices target the drivers' welfare, as they are the main actors responsible for the vehicles within the fleet and the cargo being transported. Currently, telematics is one of the leading technologies in fleet management, where data acquisition is performed by tracking and diagnostics devices. Telematics enable capturing large amounts of driving and vehicle data; this information assists companies to identify, understand and define strategies for incident prevention. Actions such as driving performance scoring, risk assessment of drivers, education and alerts are now widely adopted. To define and to follow such procedures, an in depth understanding of driving behaviour and their responses to the environment is presented in Section 2.5.2.

In this section, we investigate driving behaviour in the UK by analysing a large telematics dataset. Our review of the literature indicates that most studies in the area are limited by the number of drivers, which is in most cases fewer than 100. Furthermore, the driving data employed previously is collected for a small number of journeys per driver; and to the best of our knowledge, there is very little literature on HGV driver profiling for the UK. We address these gaps by analysing a much larger dataset, in which hundreds of thousands of HGV driving incidents in most roads in the UK for the year of 2015 are considered. In total, 21,193 HGV drivers are investigated. We want to answer the following research questions: (i) What are the existing patterns of behaviour within the UK HGV driving community? (ii) How do these patterns complement the current knowledge in the literature? To answer these questions we apply a two-stage classification framework to the dataset.

4.3.3 Dataset Description

The dataset employed in this case study is obtained in collaboration with a telematics provider company, namely, Microlise [154]. All data generated by the telematics systems were transmitted and collected from vehicles in real-time across the year [60]. The lorries are equipped with standard sensors as specified in the Society of Automotive Engineers (SAE) standard J1939-71. Data was gathered by a collection of sensors connected to multiple electronic control units by controller area network (CAN)-bus. The data was subsequently

transferred to the telematics unit that populates Microlise's databases via 3G. Microlise's telematics solutions are implemented in roughly 25% of the HGV vehicles across the UK.

For analysis in this work, 21,193 anonymised drivers were considered. Further details on how the data has been merged is found in [60]. The requirement for a HGV driver to be included in the dataset is that they must have taken part in at least 10 journeys per quarter of the year. Each driver therefore has a minimum set of 40 journeys travelled on any road of the UK. The period considered is between the first of January until the thirty first of December of 2015. Driving style data includes measures such as engine revs (to identify green band driving and over-revs), fuel consumption, accelerator position, idling, use of cruise control, use of power take off, use of primary and secondary (engine) braking, harsh acceleration and harsh braking. Many of these features collected, however, are not present in all vehicles and are therefore not available to all drivers. To provide a consistent, standardised analysis, we select the following attributes, which are present in all vehicles:

1. Driving time in seconds
2. Average daily distance in metres
3. Number of harsh braking events
4. Over speed duration in seconds
5. Excessive throttle duration in seconds
6. Number of over rev events

In addition, the number of journeys per quarter is necessary to determine those drivers to be included. Before analysis, the dataset was normalised by dividing the number of harsh brakes, overspeed, excessive throttle and over revs per driving time to avoid any biased results.

This dataset has been previously studied in the Driver of the Year competition, promoted by Microlise. The driver ranking based on incidents for the competition was performed by Figueredo *et al.* [60]. The idea of determining drivers' profiles as a continuation of the previous data exploratory work came from the necessity of the company to better understand how HGV drivers in the UK behave in order to develop better policies, and also due to the scarcity of related literature. From the analysis for the competition, three clusters based on the total distance travelled (short, medium and long distance) were identified. This study extended the authors' work by sub-clustering drivers based on their average daily distance

travelled, and categorises the drivers in groups using the two-stage classification framework, detailed in the next sub-section.

4.3.4 Classification framework tailored to Heavy Goods Vehicle dataset

Tseng *et al.* [155] hypothesise that 'the driving mileage is a determinant factor in driver speeding incidents'. The goal of this work is to determine whether (and how) distance and time driven possibly affect the occurrence of incidents in the UK. The average distance driven per day is therefore adapted instead of total distance, as it seems better suited to the driving behaviour. Once the distance ranges are identified, the next phase is to identify driving profiles within them. As mentioned previously, the analysis in this section is conducted in two stages. The framework, as tailored to the HGV problem is shown in Figure 4.4. In the first stage, consensus clustering is performed to elucidate core groups, detailed as follows:

STAGE 1

Data characterisation and pre-processing

If necessary the data is cleaned, normalised and inconsistencies are fixed. For our dataset we perform consensus clustering employing the attributes harsh braking events, over speed duration, number of excessive throttle events and number of over rev events. The features across the dataset are normalised by the driver's total driving time in seconds. Figure 4.5 shows the exploratory analysis of the data. The data are skewed as drivers with high numbers of incidents are rarer (as expected). We have, however, not removed what may appear to be outliers in the box plots distributions shown below, as extreme behaviour and the sparsity of the data instances are important for driving characterisation. Further discussion on the importance of keeping what appear to be outliers in the analysis is given in the next sections, as we conduct the analysis.

No further transformation has been applied to the dataset as we want to observe how the grouping of the real-world behaviour occurs, i.e., to establish a clear separation between bad drivers and good drivers. Note, we have also attempted to cluster the data after a min-max normalisation, but no relevant cluster is obtained from this exercise, therefore they were not included in the chapter.

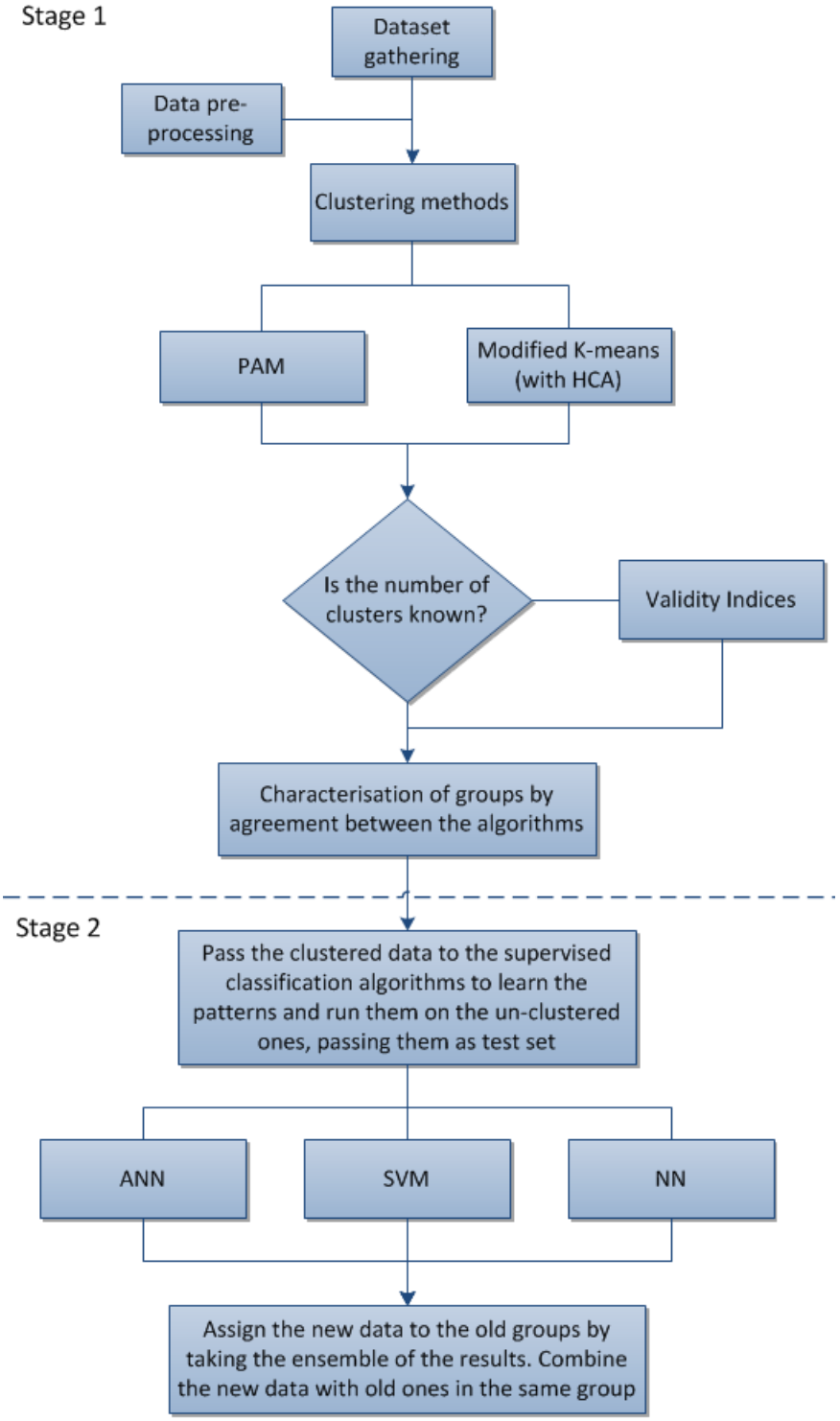


Figure 4.4 Flowchart of the framework to determine the clusters in the drivers data set

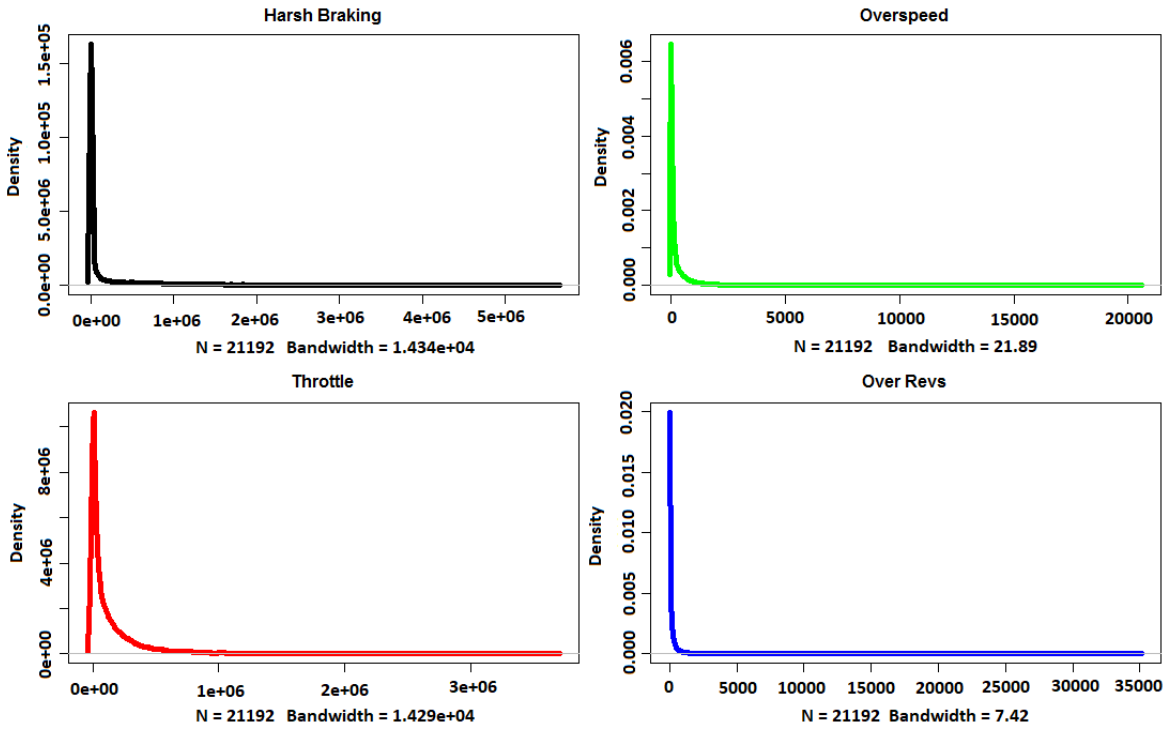


Figure 4.5 Data set features density distribution

Clustering

The K-means with candidate cluster centres from HCA and Partitioning Around Medoids were the two clustering employed, similar to the previous case studies. These techniques differ from each other as they have different measurements to define proximity of the data instances to establish the clusters. This means that they group the data considering different characteristics.

Determining the number of clusters

In this step, validity indices were applied to the clustering results. As for the HGV problem the number of clusters are unknown, therefore in an attempt to determine an ideal value, the indices Calinski and Harabasz, Scott and Symons, Marriot and Trace W were applied to K-Means and PAM, for which the number of clusters and cluster centres are explicit parameters, over a range of number of clusters varying from 2 to 20. According to specific rules, validity indices indicate the appropriate number of groups to consider in the analysis. In the indices implementation, cluster stability is also assessed.

Consensus

Consensus between the algorithms is achieved by aligning the clusters found by the different techniques. The agreement among solutions of the different clustering methods employed is evaluated using the Cohen's kappa coefficient k [156]. This coefficient is a statistical measure of agreement for categorical items.

STAGE 2

After employing the consensus clustering, the instances without a common consensus on the final class label are categorised as *unclassified*. In order to reduce the number of unclassified instances, the second stage of the framework is employed, described below:

Ensemble Classification

The data points which were classified in Stage 1 are used to train the ensemble classification models and the remaining unclassified data are used as test set. We attempt to classify the previously unclustered data (test set) with the goal to assign them to one of the previously identified groups. Support Vector Machines (SVM) [36], Nearest Neighbour (NN) [157] and Multi-Layer Perceptron Neural Networks (MLP) [158] are employed as part of the ensemble to learn the patterns. Among a number of decision level fusion methods present in the literature, majority voting is employed in this case study.

To evaluate and validate the characteristics of the data assigned to clusters from Stage 1 after ensemble classification, visual tests, statistical tests and cluster quality assessment need to be performed. The distribution of each attribute in a cluster after Stage 1 is compared to that from Ensemble Classification from Stage 2 using boxplots, the Mann-Whitney-Wilcoxon non-parametric t-test at 0.05 level of significance. If a group of newly classified instances is statistically similar to any of the previous clusters after the tests, the drivers in the newly classified group are combined with the equivalent group of drivers previously clustered after Stage 1. Otherwise, the group of misclassified instances (instances classified as belonging to a certain cluster by the ensemble, but the statistical tests fail to confirm the classification) is checked for the possibility of being an extra profile. Lastly, the majority voting ensemble classification results are evaluated using the Davies-Bound (DB) internal cluster quality assessment index. The DB index computes the ratio of intra- and inter-cluster distance of the drivers. Therefore, to study the quality of clusters, the DB index values of each subgroup of drivers assigned after ensemble classification are compared to the values of the DB index obtained if they had been assigned to each of the other clusters.

Algorithm 1: Label behaviour

inputs : Box-plot for $variable_i$ (i:1..4) in cluster, box-plot for all data for $variable_i$ for subgroup.

output : $label_i$ of behaviour for each incident variable

foreach *Incident variable* v_i, *i:1..4* **do**

  if $Median(variable_i) \leq Median(AllDataInSubgroupForVariable_i)$ **then**

    $label_i \leftarrow$ "Low";

  else if $Median(variable_i)$ *IsBetween*$(Median(AllDataSubgroupForVariable_i)$ *and* $ThirdQuartile((AllDataSubgroupForVariable_i)))$ **then**

    $label_i \leftarrow$ "Moderate";

  else if $Median(variable_i)$ *IsBetween* $ThirdQuartile(AllDataInSubgroupForVariable_i)$ *and* $Maximum(AllDataInSubgroupForVariable_i)$ **then**

    $label_i \leftarrow$ "High";

  else if $Median(variable_i) > Maximum(AllDataSubgroupForVariable_i)$ **then**

    $label_i \leftarrow$ "Very High";

After Stage 1 is applied to 6,600 instances, 3 clusters are found, with 600 instances as unclassified. 6,000 data points from clusters 1, 2 and 3 are used to train the classifiers at Stage 2. The 600 unclassified instances are subsequently classified by models within the ensemble using majority vote. Let us assume that the result of the 600 instances classification is as follows: 100 instances are classified as belonging to cluster 1 (from Stage 1); 200 instances are classified as part of cluster 2; and 300 instances are classified as belonging to cluster 3. To ensure that this classification is correct, we subsequently resort to the box-plots, Mann-Whitney-Wilcoxson test and the DB cluster quality assessment. Therefore,we plot the distribution of the original cluster 1 (from Stage 1) and contrast with the distribution of the 100 instances newly classified; we also compare them using the statistical tests. If those tests show that there is no visual or statistically significant difference between cluster 1 and newly classified 100 instances, then these 100 instances are merged into cluster 1. Otherwise, they remain unclassified or, if the group is large enough, their instances are tested to assess if they represent a new cluster (or pattern of driving behaviour). The same process is repeated for the other clusters.

After ensemble classification, more data are expected to be assigned to one of the groups, and the remaining data are still unlabelled or 'unclassified'. The clusters obtained from Stage 1 are combined with the newly classified data in the same groups to achieve the final clusters.

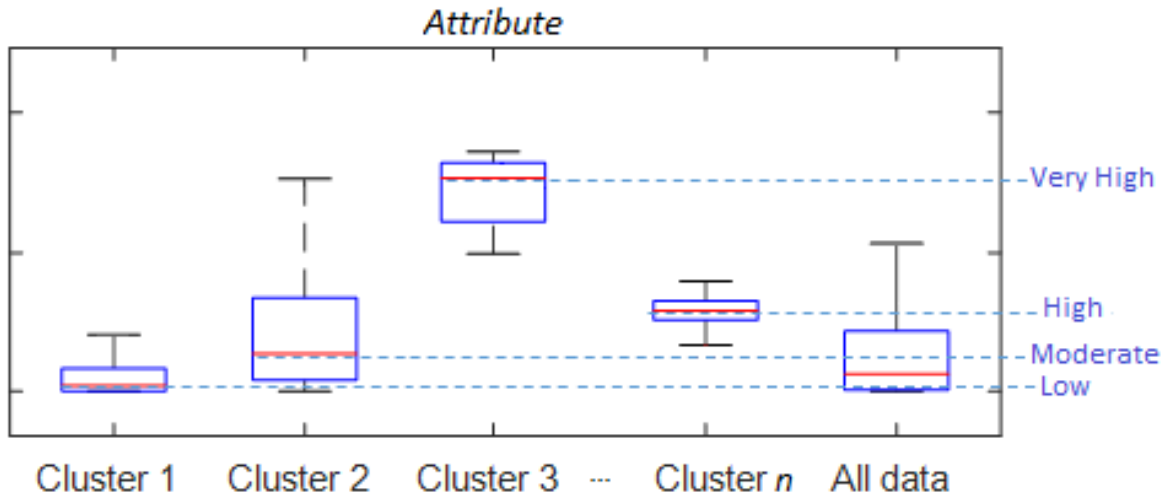


Figure 4.6 Example of the comparison of each variable with the median values found in the cluster. The corresponding labels are on the right side of the figure.

4.3.5 Driver Profile Labelling Algorithm

To establish driver profiles (driving stereotypes), we put forward a novel profile labelling algorithm that compares the median values for each variable of each cluster (i.e., harsh braking, over speed, excessive throttle and over revs) with the quartiles in the box-plot for the entire group data. The categories defined for the attributes are **low**, **moderate**, **high** and **very high** occurrence of incidents, as detailed in Algorithm 1. Figure 4.6 shows the labels of boxplots used to define the profiles.

In the next subsections the experiment and results of the framework, as applied to the HGV data is presented.

4.3.6 Experimental Settings

Two instances of SVM classification algorithm are used in the ensemble, with radial basis function and sigma set at $0.5/n_f$ and $1/n_f$ respectively, where n_f is the number of features in the dataset. In addition, two MLP classification models are considered with 1 and 3 neighbours with Euclidean distance function. MLP is adopted with three hidden layers and a sigmoid activation function to model the weighted sum. Three instances of MLP are part of the ensemble: (i) 15 nodes for layer 1, 20 nodes for layer 2 and 15 nodes for layer 3 with 100 epochs; (ii) 10 nodes for layer 1, 15 nodes for layer 2 and 10 nodes for layer 3 with 100 epochs; and (iii) 10 nodes for layer 1, 20 nodes for layer 2 and 10 nodes for layer 3 with 100 epochs. All algorithms are implemented using 10-fold cross validation.

Table 4.7 Three main subgroups based on average distance travelled

Groups	Average daily Km range	Number of drivers
1 - Low daily average	from 36Km to 220Km	5077
2 - Medium daily average	from 220Km to 350Km	8391
3 - High daily average	from 350Km to 750Km	7725

4.3.7 Results

In Figueredo *et al.* [60], three subgroups based on distance are identified (short, medium and long distance drivers). We adapt these groups to consider the average distance driven per day. K-Means grouped the distance ranges and the proportions for the low, medium and high daily average distance groups, as shown in Table 4.7. The two-stage framework is therefore applied to the data subsets within high, medium and low daily average distance travelled. For the next phase, the consensus clustering is run on the subgroups found.

Low Daily Average Distance Travelled Subgroup

Validity indexes results indicate that three and four clusters are the most suitable numbers (see Figure 4.7). Kappa index of 0.7725 on the consensus results confirms four clusters should be adopted (Kappa for 3 clusters is 0.3757). Stage 1 results for four clusters are shown in Table 4.8 (second column). 4,471 out of 5,077 drivers within this group are assigned to one of the identified clusters; 606 however are associated with no group. Stage 2 is therefore applied to the unclassified data; results are presented in the third column of Table 4.8. It contains the number of unclassified drivers re-assigned to clusters, with p-value confidence interval 0.05. These results were also verified by the DB index. The 'Total' column in Table 4.8 shows the final classification of drivers after combining the results of Stage 1 and 2. The number of unclassified drivers is reduced to 217.

Ensemble classification results in Stage 2 show that 148 instances have been assigned to cluster 3 and another 241 instances form a new cluster. We achieved this outcome by employing the classification ensemble to the unclassified data and performing a statistical comparison on the results. Table 4.9 shows the p-value comparison (for each feature) of the existing clusters found in Stage 1 against the new clusters formed after the ensemble classification (Stage 2). Both unanimous and majority voting ensemble algorithms are employed. In this particular case of low daily average distance drivers, both unanimous and majority voting show similar results. To achieve maximum classification, however, the drivers assigned to clusters in Stage 2 are obtained after majority voting (Table 4.9).

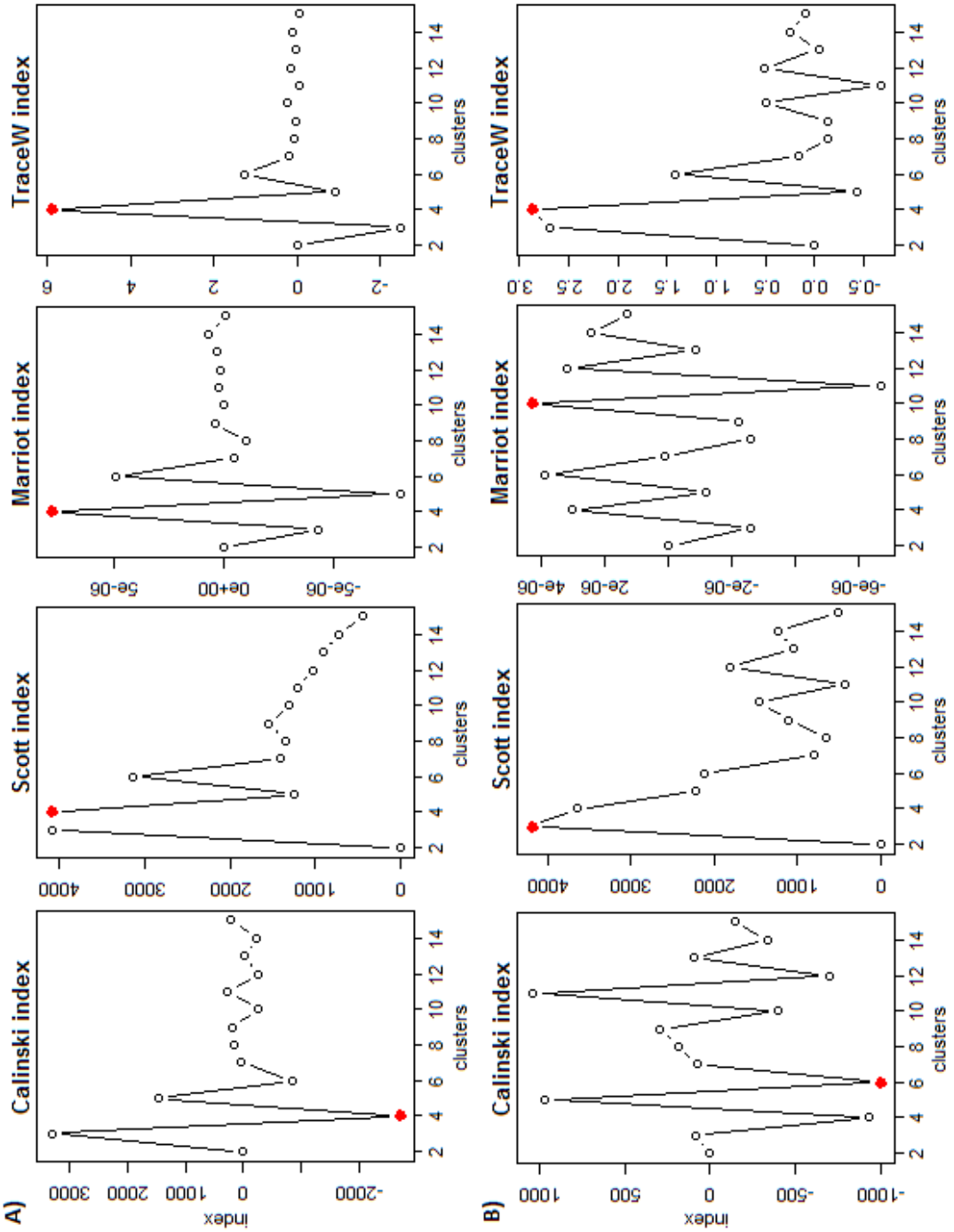


Figure 4.7 Validity indices for (A) K-means and (B) PAM for the instances with low daily average mileage

Table 4.8 Distribution of drivers among the groups for Low Daily Average Distance

Cluster	Stage 1	Stage 2	Total
1	3,107	-	3,107
2	177	-	177
3	432	148	580
4	755	-	755
New Cluster	**-**	**241**	241
Unclassified	606	217	217

Table 4.9 p-value of majority voting for Low Daily Average Distance

Driving Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Harsh Braking	0.08	0.004	**0.15**	0.73
Overspeed	$2.2x10^{-16}$	$2.2x10^{-16}$	**0.02**	0.5
Throttle	$2.2x10^{-16}$	0.008	**0.11**	$2.2x10^{-16}$
Over Revs	0.27	0.46	**0.22**	0.04
Number of drivers	**241**	60	148	157

The instances assigned to cluster 3 (148 instances) have p-values greater than 0.05, except for overspeed. The statistical test results confirm that those 148 instances belong to cluster 3 found previously in Stage 1. The drivers designated to the other three groups by the ensemble classification had p-value smaller than 0.05 for most attributes, which indicates that they do not belong to the existing clusters and they are misclassified.

It is interesting to note at the bottom of the majority vote Table (4.9), that after employing the complete classification framework 241 misclassified drivers are from cluster 1; 60 misclassified drivers belong to cluster 2 and 157 misclassified being part of cluster 4. To check whether these groups of instances can possibly form a new pattern of behaviour, Algorithm 1 is run on the driver groups of 241 drivers from Stage 2, as they form a large group. Results are compared with the current profiles to check if there are any differences in behaviour. As a consequence, a new driving profile is uncovered. The remaining non-assigned instances are discarded (continue being labelled as 'unclassified'), as they are low in number for a new profile.

The driving patterns of behaviour are shown in Table 4.10. Figure 4.8 shows the cluster characterisation of these groups using box plots. Figure 4.9 shows the values for each variable in the clusters against the median of all low average daily distance drivers. From the figures, it is possible to observe that in the first cluster, the number of incidents are low. The second cluster comprises of drivers with the very high over speeding and high throttle values. Cluster 3 is characterised by high over speed and throttle, although these values are lower than in

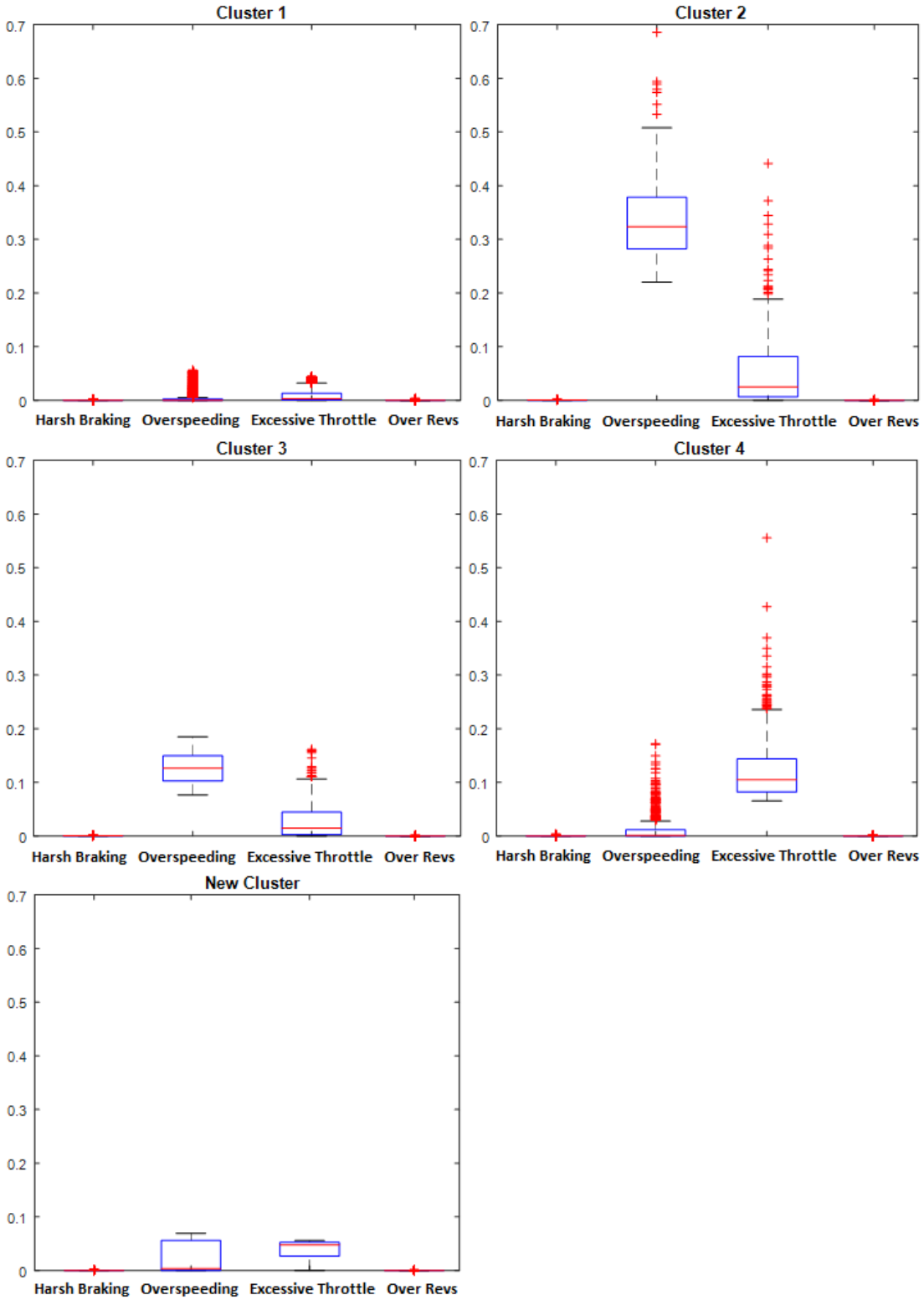


Figure 4.8 Clusters within low daily average distance

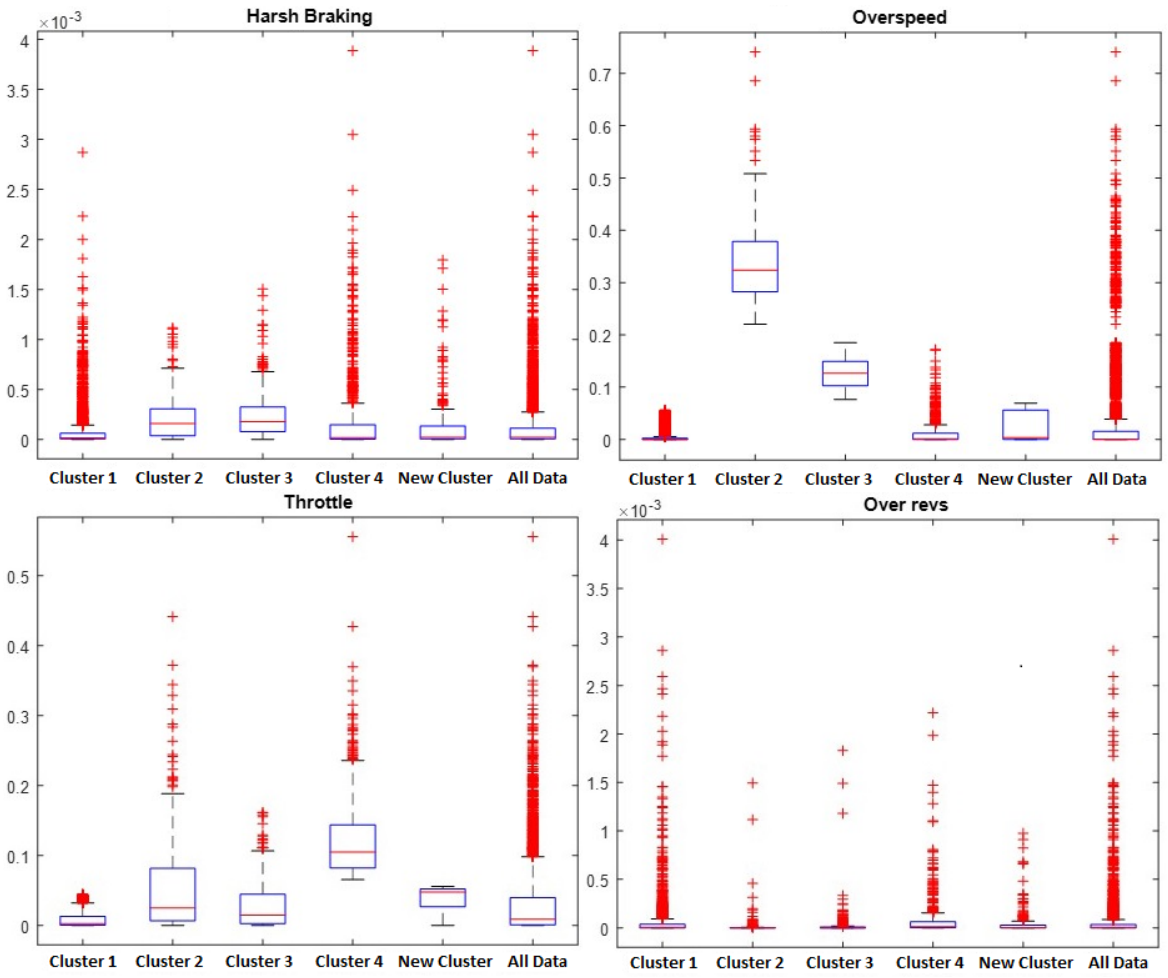


Figure 4.9 Clusters within low daily average distance: comparing each variable with the median values found in the cluster.

cluster 2. Cluster 4 has an overall high throttle. The new cluster from Stage 2 has moderate values for harsh braking, overspeed and over revs, but has high values for throttle and low over revs. It can be observed that the algorithm proposed produces the same pattern for the second and third rows of the table. Although in Figure 4.9 there is a distinct difference between the values of overspeed between these groups, we give them the same label, as for both cases the median is above the maximum value from the box-plot of the entire dataset. In the future, however, if experts identify the need of distinguishing between both over speed behaviours, another label for speed (i.e., 'extremely high') can be added to the algorithm. Lastly, it is important to notice that the outliers in the two figures represent the extreme driving behaviours within the identified clusters, shown in detail in the discussion section.

Table 4.10 Profiles for Low Daily Average Distance

Cluster	Harsh Braking	Overspeed	Throtle	Over Revs	Profile
1	Low	Low	Low	Moderate	1
2	High	High	Moderate	Low	2
3	High	High	Moderate	Low	2
4	Low	Moderate	High	Moderate	3
New Cluster	Low	Moderate	High	Low	4

Table 4.10 shows the driving profiles after employing Algorithm 1 on the low daily average distance instances. It is possible to identify four patterns of behaviour: very safe drivers, who comply with traffic rules and have mostly moderate amounts of over revs; unsafe drivers, with high values of overspeed and harsh braking; groups 3 and 4 comprise mildly unsafe drivers, with distinct characteristics between them. Group 3 presents low harsh braking, moderate overspeed and over revs and high throttle. Group 4 has moderate values for harsh braking and overspeed, high throttle and low over revs.

Medium Daily Average Distance Travelled Subgroup

Validity indexes suggest the data should be divided either in three (second best number in certain cases, see Figure 4.10), four or six clusters; however, three clusters produce a higher kappa (0.8026) and a smaller number of unclassified instances (888). 7,503 out of 8,391 drivers are distributed in three groups after Stage 1 (Table 4.11, second column). Stage 2 adds 288 drivers to the second cluster and detects a new cluster with 502 drivers (Table 4.11, third column). The final number of unclassified instances is reduced to 98. The p-value comparison of clusters detected after Stage 1 with those from Stage 2 is shown in Table 4.12. This classification was also confirmed by the DB index.

Figures 4.11 and 4.12 show the cluster results for the medium daily average distance dataset. The first cluster comprises a pattern of behaviour identified previously, as it contains the group of safe drivers, with low values for all incidents. Cluster two differs from the first

Table 4.11 Distribution of drivers among the groups for Medium Daily Average Distance

Cluster	Stage 1	Stage 2	Total
1	4,863	-	4,863
2	892	288	1,180
3	1748	-	1,748
New Cluster	-	502	502
Unclassified	888	98	

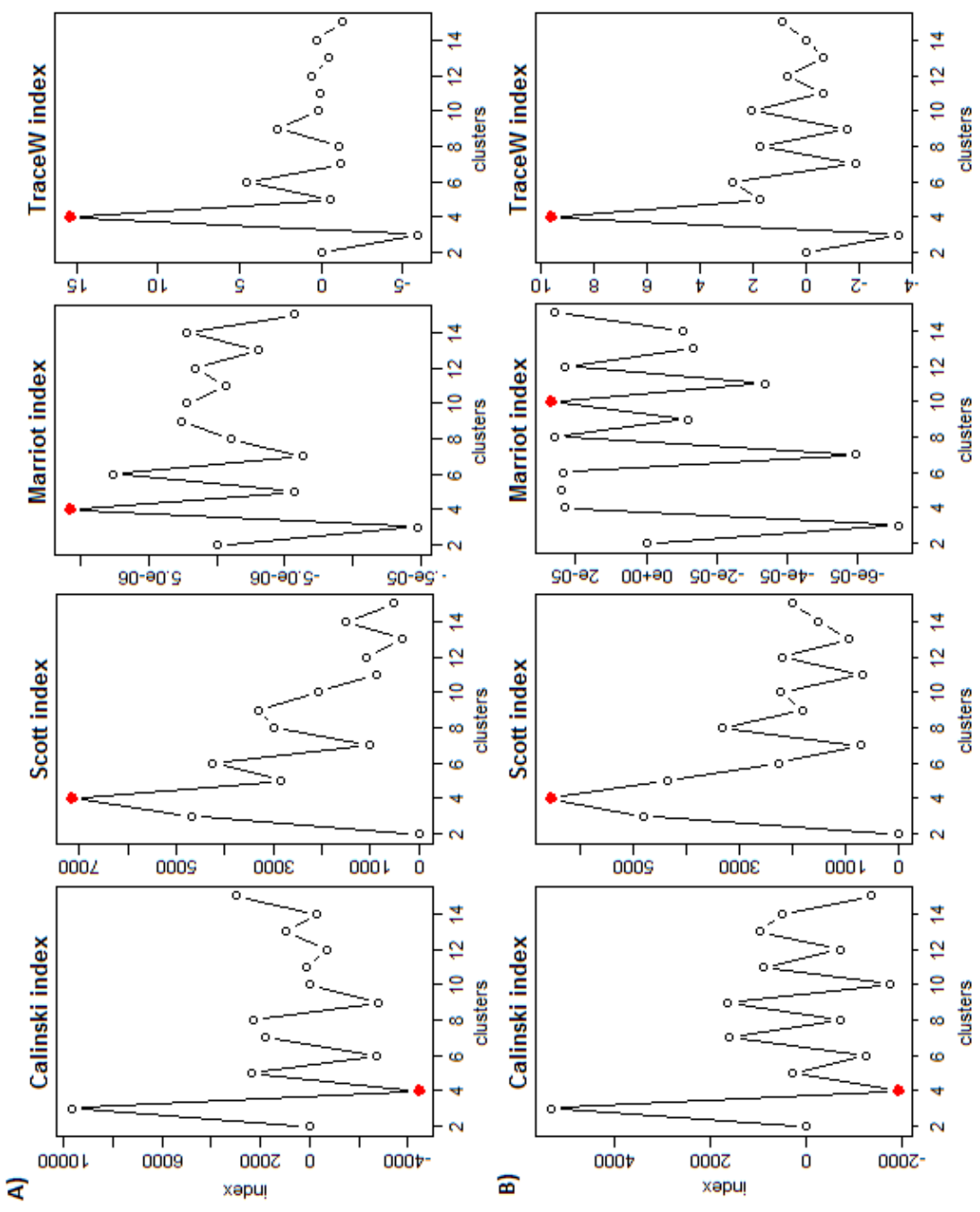


Figure 4.10 Validity indices for (A) K-means and (B) PAM for the instances with medium daily average mileage

Table 4.12 p-value comparison of ensemble clustering with ensemble classification for Medium Daily Average Distance

Driving Feature	Cluster 1	Cluster 2	Cluster 3
Harsh Braking	$6.1x10^{-6}$	**0.4**	0.07
Overspeed	$2.2x10^{-16}$	**0.75**	$2.2x10^{-16}$
Throttle	$2.2x10^{-16}$	**$2.2x10^{-16}$**	0.02
Over Revs	$7.2x10^{-4}$	**0.41**	$8.1x10^{-5}$
Number of drivers	502	288	98

Table 4.13 Profiles for Medium Daily Average Distance

Cluster	Harsh Braking	Overspeed	Throtle	Over Revs	Profile
1	Low	Low	Low	Moderate	1
2	Low	Low	High	Moderate	5
3	High	High	Moderate	Low	2
New Cluster	Low	Moderate	High	Low	4

cluster mostly by the number of excessive throttle, which is much higher. The graph for cluster three presents high values for overspeed and throttle, and the fourth cluster appears to have lower values for harsh braking and over revs and higher values for overspeed and throttle. Table 4.13 shows the profiles detected after employing Algorithm 1 on medium daily average distance drivers. Among the four driving profiles, two are also present in the low average group, as indicated in the last column of the table.

The difference in p-values for new groups is more noticeable in this case (Table 4.12). Algorithm 1 confirms one new driving profile assigning 502 drivers to them, represented by 'New Cluster' in Table 4.11. The Total column in Table 4.11 combines the drivers from Stage 1 and Stage 2 for final classification. After Stage 2, the number of unclassified drivers was massively reduced from 888 down to 98.

High Daily Average Distance Travelled Subgroup

For the last subgroup, although there is more consensus (less unclassified instances) with three clusters (905 unclassified and kappa equals 0.7658), the kappa coefficient favours four clusters (0.7776). In addition, more patterns of behaviour are uncovered when four clusters are considered. After Stage 1, 6,810 out of 7,725 drivers are distributed in four clusters (Table 4.14, column 2). The third column shows Stage 2 results, in which 31 instances are assigned to cluster 2; 254 drivers are added to cluster 3 and 59 are associated with the fourth cluster. A new cluster is identified and contains 569 drivers. The number of unclassified

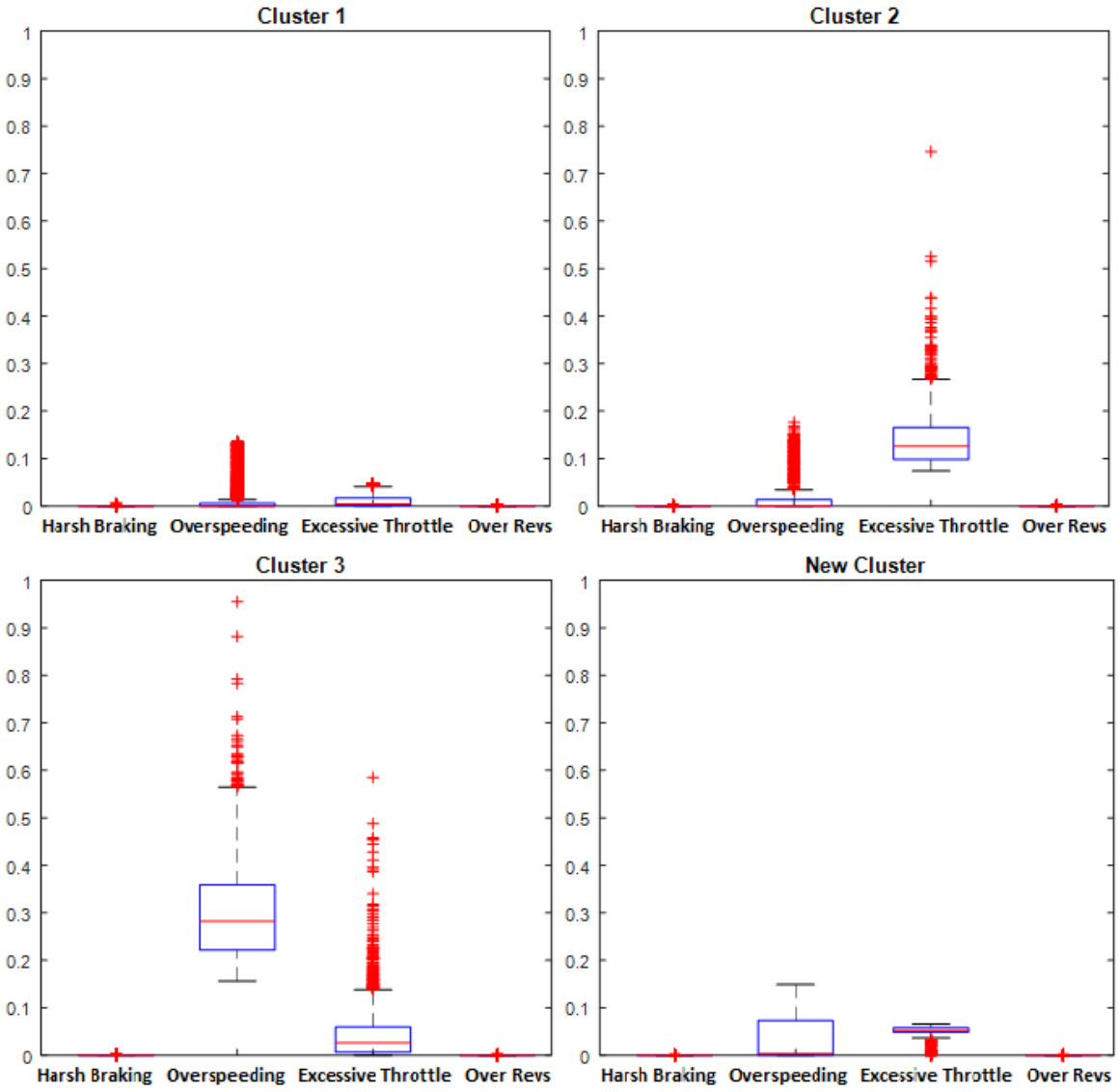


Figure 4.11 Clusters within medium daily average distance

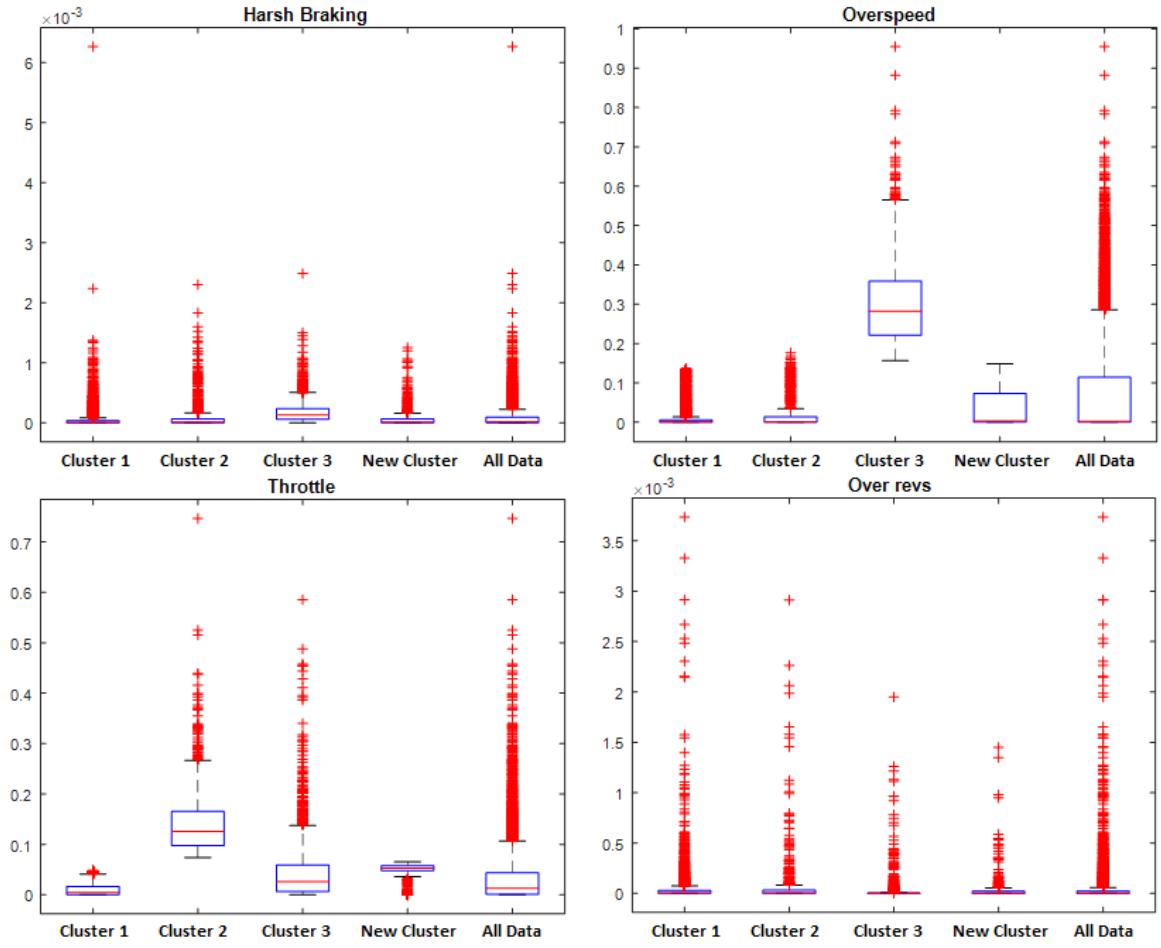


Figure 4.12 Clusters within medium daily average distance: comparing each variable with the median values found in the cluster.

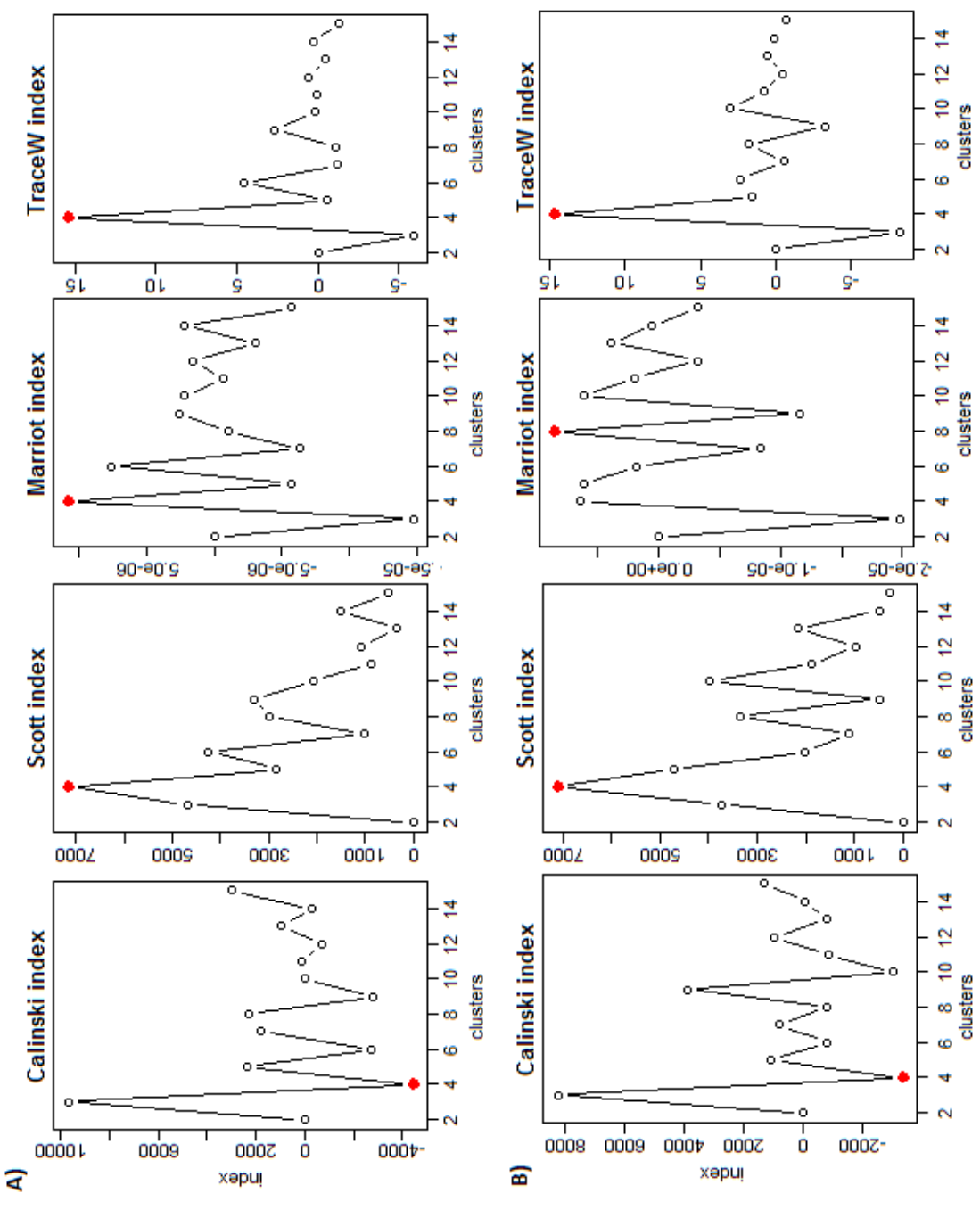


Figure 4.13 Validity indices for (A) K-means and (B) PAM for the instances with high daily average mileage

Table 4.14 Distribution of drivers among the groups for High Daily Average Distance

Cluster	Stage 1	Stage 2	Total
1	4,657	-	4,657
2	1,007	59	1,066
3	604	254	858
4	542	31	573
New Cluster	-	569	569
Unclassified	915	2	

Table 4.15 p-value comparison of ensemble clustering with ensemble classification for High Daily Average Distance

Driving Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Harsh Braking	$8.7x10^{-8}$	**0.71**	**0.87**	**0.08**
Overspeed	$1.1x10^{-13}$	$\mathbf{7.1x10^{-12}}$	**0.06**	$\mathbf{2.2x10^{-16}}$
Throttle	$2.2x10^{-16}$	**0.7**	$\mathbf{2.2x10^{-16}}$	**0.52**
Over Revs	$3x10^{-15}$	**0.71**	**0.48**	**0.18**
Number of drivers	569	59	254	31

Table 4.16 Profiles for High Daily Average Distance

Cluster	Harsh Braking	Overspeed	Throttle	Over Revs	Profile
1	Low	Low	Low	Moderate	1
2	High	High	Moderate	Low	2
3	Moderate	Moderate	High	Low	6
4	Moderate	High	High	Low	7
New Cluster	Low	Low	High	Low	8

drivers is reduced to 2. The p-value comparison of clusters detected after Stage 1 with those from Stage 2 is shown in Table 4.15, once again verified by DB index.

Figures 4.14 and 4.15 characterise the clusters identified. In the table there is a cluster with low values of incidents (cluster one); cluster two is characterised by high harsh braking and overspeed and moderate throttle, whereas cluster three has drivers with high values for harsh braking and throttle. The fourth cluster is mostly characterised by high overspeed. The new cluster has high throttle but presents low values for all the other variables. The complete characterisation of the stereotypes is found in Table 4.16. In the table, three new driving styles are found. These patterns are not identified for both low and medium daily average distance driven.

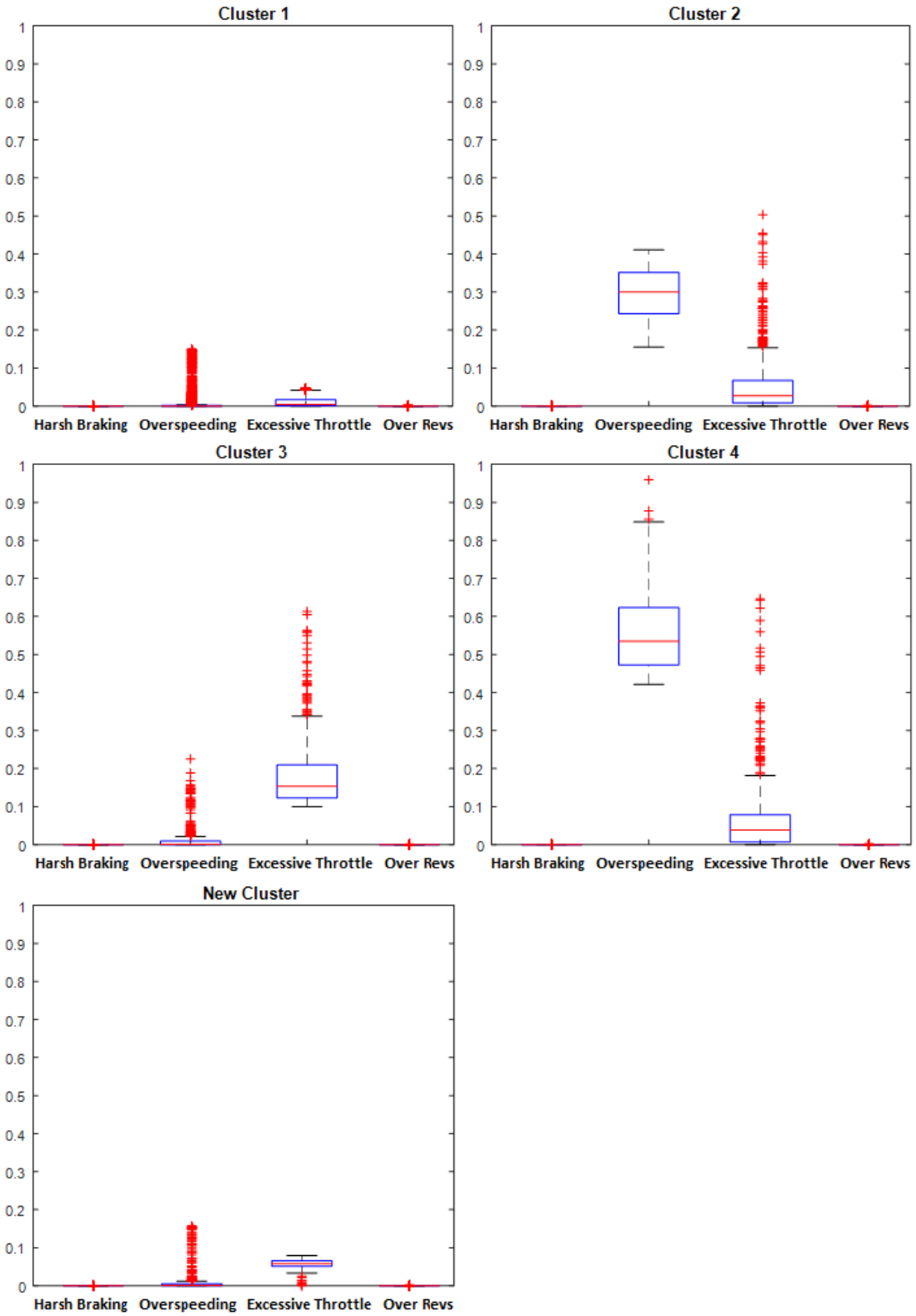


Figure 4.14 Clusters within high daily average distance

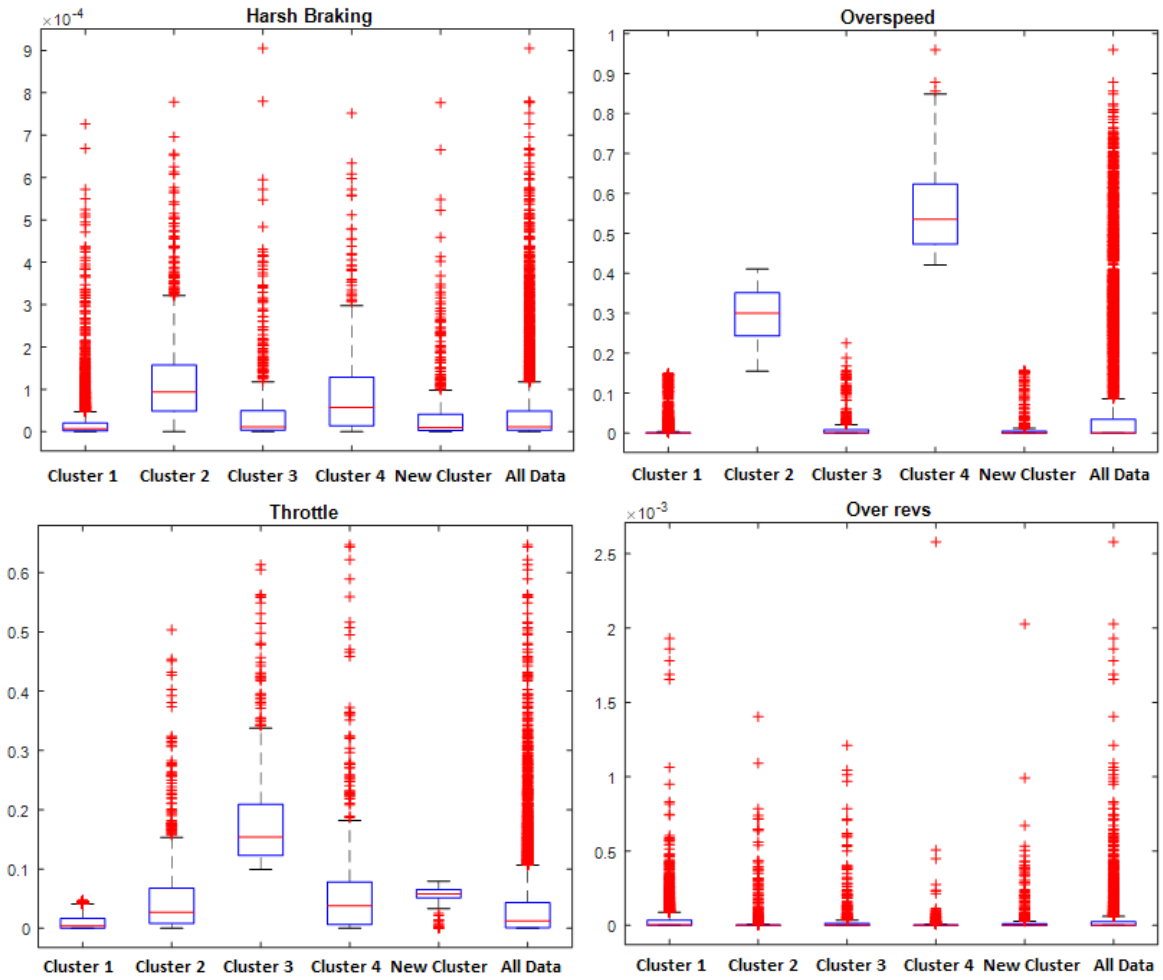


Figure 4.15 Clusters within high daily average distance: comparing each variable with the median values found in the cluster.

4.3.8 Discussion

Profiles Identified

In this case study on HGV dataset, eight driving profiles were uncovered. For each stage, the final set of driving profiles is displayed in Table 4.17, along with the number of drivers present in each group and their proportions in relation to the entire dataset. The sixth column shows the numbers after the first stage; the seventh column displays the final results, after employing Stage 2. The last column shows in which subgroups (low (L), medium (M) or high (H) average daily distance) the profiles are found.

The first stereotype (Profile 1) comprises of safe drivers and is common to all the three subgroups and represents the largest proportion of drivers in the dataset (59.58%). Profile 2

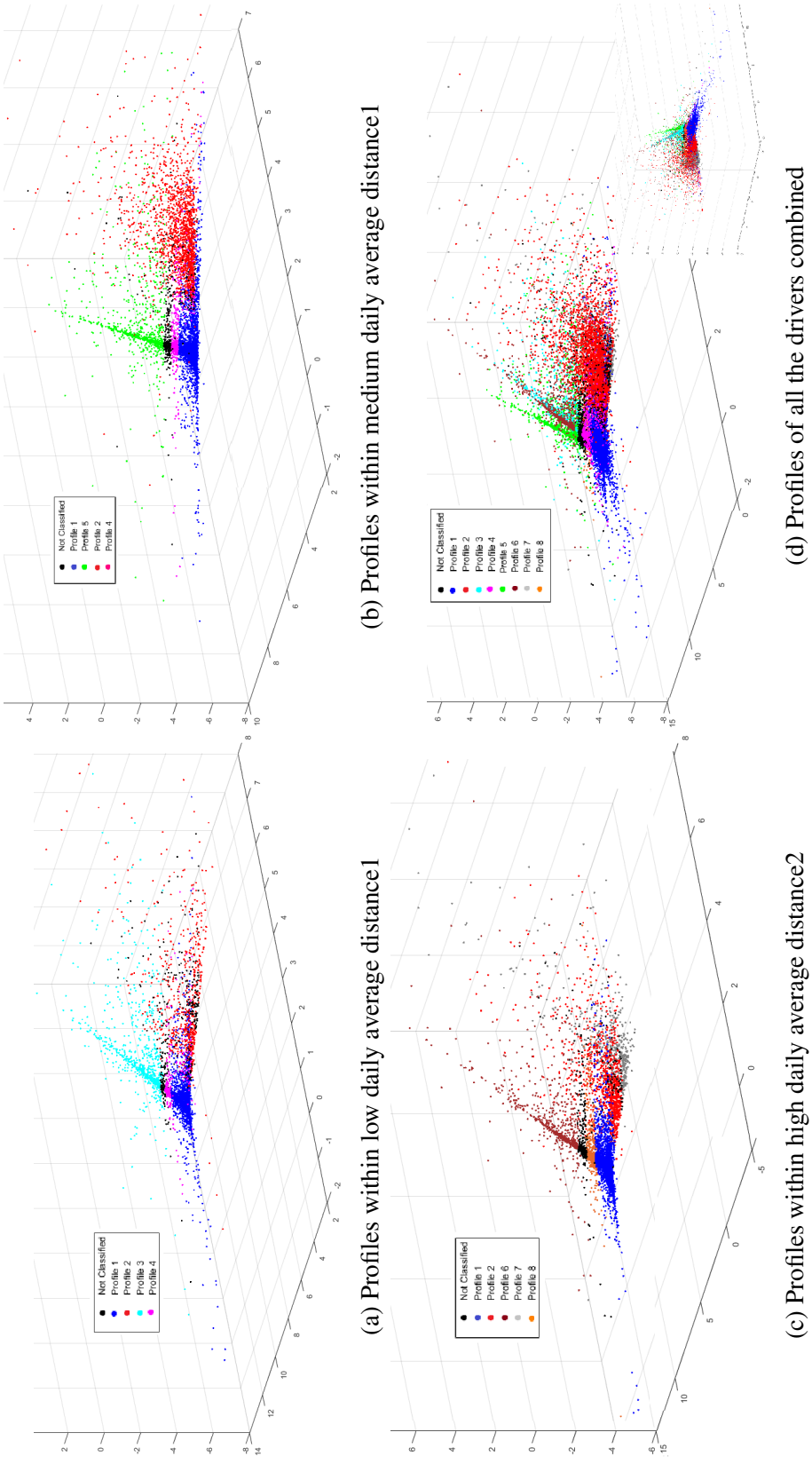
(aggressive drivers) is also common to all the three subgroups, characterised by high harsh braking and over speeding, but moderate throttle. It is the second profile highest in proportion, with 16.72%. The remaining profiles seem to appear in lower proportions within the UK drivers. Profiles 3 and 4 have a similar number of harsh braking, overspeed and excessive throttle but differ only in over revs events. Profiles 5 and 8 present low harsh braking, low over speeding and high throttle; however, they differ only with regards to the number of over revs. It is important to notice that profiles 5 and 8 comprise medium and high daily average distance drivers, respectively. Drivers in profiles 6 and 7 differ by the number of overspeed events. The high throttle for profiles 5 and 8, rather than indicating bad driving behaviour, might suggest peculiarities of roads (high inclination), or the vehicles or the nature of the cargo being transported. Further information is necessary to verify this hypothesis. If further evidence is provided, profiles 1, 5 and 8 might indicate very similar or even the same behaviour, which increases the proportions of very safe drivers.

After employing the classification framework, only 317 drivers remained unclassified. Further investigation is necessary to identify the profile to which these drivers belong. Figure 4.16 shows the 3-D scatter plot of the clusters in the three driving groups. These were obtained by taking the principal components on the three groups. Figure 4.16 (a), (b) and (c) show the drivers among 5 groups for low, 4 groups for medium and 5 groups for high average distances, respectively. Lastly, figure 4.16 (d) show all the profiles combined in one scatter plot (Note: This figure also show the scatter plot rotated by 180 degrees, to show the profiles on the reverse side of the image). The four figures represent the driving behaviours within the identified profiles among all the three subgroups. In the figures, the corresponding profiles in all graphs have the same colour. For instance, Profile 1, which is common to all subgroups of average distance is shown in blue. Points in black are the unclassified instances. Profiles that are not common for the subgroups are represented in different colours.

From the figure it is possible to observe why a number points that seem to be outliers, appear in the box plots of the clusters. There are a number of instances which are more sparse compared to the other points in the clusters, even though these instances still belong to the same pattern of behaviour as the clusters they have been assigned to. These points, however, are not outliers. The consensus clustering method would have separated outliers into extra clusters. In addition, we have attempted to remove the outliers that appear in the box-plots, and as a consequence, new outliers appear, demonstrating that the data is spread within the clusters. We believe that within a specific stereotype, these instances represent the extreme values.

Table 4.17 Driving Profiles

Profile Number	N. of Harsh Braking Events	Overspeed	Excessive Throttle	N. of Over Revs Events	N. of Drivers Classified after Stage 1	N. of Drivers classified after Stage 2	Sub-Group Label
1	Low	Low	Low	Moderate	12,627 (59.58%)	12,627 (59.58%)	L,M,H
2	High	High	Moderate	Low	3,364 (15.87%)	3,543 (16.72%)	L,M,H
3	Low	Moderate	High	Moderate	755 (3.56%)	755 (3.56%)	L
4	Low	Moderate	High	Low	–	743 (3.51%)	L,M
5	Low	Low	High	Moderate	892 (4.21%)	1180 (5.57%)	M
6	Moderate	Moderate	High	Low	604 (2.85%)	663 (3.13%)	H
7	Moderate	High	High	Low	542 (2.56%)	796 (3.76%)	H
8	Low	Low	High	Low	–	569 (2.69%)	H
Unclassified					2,409 (11.37%)	317 (1.5%)	



(a) Profiles within low daily average distance1

(b) Profiles within medium daily average distance1

(c) Profiles within high daily average distance2

(d) Profiles of all the drivers combined

Figure 4.16 Scatter plot of the drivers in their respective driving profiles for (a) low, (b) medium (c) high daily average distance and (d) the entire data set

Comparison with the State-of-the-Art

Quantitative comparison between the results of this case study and results in literature is challenging due to differences in the data collected techniques, the types of drivers included and the methodology employed. Although, this case study has confirmed the existing research from literature [12], as few previously identified stereotypes were also found in our analysis. For instance, the profiles identified by Constantinescu *et al.* [12] (in Chapter 2 Table 2.2) is loosely similar (where the attributes considered in the data are the same) to the profiles of well-behaved drivers (i.e. 1, 5 and 8). Similarly, profiles 2 and 7 from this study match their sixth cluster. The other groups found would require further investigation, as their criteria for the classification of profiles and the attributes used for clustering are different. Other authors, such as Halim *et al.* [122] have also identified those types of behaviour for smaller vehicles.

Qualitatively comparing the results of the framework with the methods found in literature, the two-stage classification frameworks is able to identify more stable clusters which are obtained as result of consensus of multiple methods. The current literature mostly employs fewer clustering algorithms and does not consider consensus. In addition, the classification frameworks allows a more detailed report on less frequent behaviour, which is statistically different from those groups detected in Stage 1. Moreover, no artificial data (produced via simulation) is used in this case study and we do not interfere on driving behaviour during the data collection.

Applications

Informing the stakeholders (managers, policy makers, vehicle manufacturers, etc.) about the patterns of behaviour and their proportions will allow them to (i) understand what are the characteristics of UK HGV drivers and (ii) to investigate, together with human factors specialists, what mitigation actions could be taken to educate drivers and to reduce the number of road incidents. In addition, for telematics companies such as Microlise, an implementation of a classification system able to predict the driver's behaviour could assist in the development of internal policies, methods for scoring, the creation of games and rewards for improved behaviour. One example of an existing action to reward driving behaviour is the Microlise Driver of the Year Competition [60]. This competition, however, focuses on past behaviour. By establishing an on-line, real-time classification approach, more effective and immediate reinforcement of good behaviour can be employed. Similarly, such classification can be used in inbuilt systems in smart cars to assess and trigger profile-specific actions for incident prevention. The profiles found can also be translated into stereotypes defining agents

behaviour in agent-based modelling simulation exercises. This allows for the creation of artificial laboratories to investigate traffic, incident prevention, responses to actions, etc.

4.3.9 Summary

In this case study, eight types of driving behaviour amongst UK HGV drivers were uncovered. To achieve these results, a dataset containing incidents over the year of 2015 for 21,193 drivers is examined. The incidents considered are harsh braking, over speeding, excessive throttle and over revving. Eight driving patterns are uncovered by employing the two stage framework, not described previously. From Table 4.17 it can be observed that after Stage 1, 18,784 drivers out of 21,193 were distributed among six driving profiles, while 2,409 (11.37%) drivers were left unclassified. Adding Stage 2, two extra driving profiles were uncovered and the number of unclassified drivers was reduced to 317 (1.5%).

These findings can assist policy makers and stakeholders in the HGV industry to better understand their drivers and to define actions to improve road economy and safety. The use of this 2-stage framework has lead to the formation of robust groups, validated by experts in the HGV industry. A real-world dataset is used in this case study, again indicating robustness and use of the novel classification framework.

4.4 Overall Summary

In the domain of unlabelled datasets, classification of instances which are difficult to classify is a challenging task. Current methods employed across unlabelled datasets face this challenge, leading to an incomplete representation of the dataset. Thus, to improve data distribution in the identified groups, a novel two stage classification framework was introduced in Chapter 3. This chapter builds upon the classification framework presented in the last chapter by empirically evaluating it on three case studies.

In the first case study, the two-stage classification framework is employed on the Edinburgh breast cancer datasets for biological group stratification. Multiple tests including Mann-Whitney-Wilcoxson, boxplots and DB index were used to evaluate the classification obtained by this framework. The results suggest that the characteristics of the obtained biological groups were similar to the corresponding ones in the literature, validating the classification performance.

The second study makes use of multiple standard datasets from UCI repository for stratification of their characteristic groups. The classification obtained by the framework

is verified by comparing its results with the ground truth was available for these datasets. Again, a good classification performance was observed, indicating the generality, robustness and effectiveness of the two-stage classification framework.

In the final case study, the framework is used to identify the existing driving stereotypes in the Heavy Goods Vehicle community within U.K. Eight state-of-the-art driving stereotypes were established, verified by experts in this domain. Moreover, these results can also enable researchers, companies and policy makers to establish group-specific strategies in order to improve safety and reduce risks.

Overall, the two stage classification framework results in an improved classification in all the three case studies, highlighting the framework's generality and capability to adapt to different types of datasets (robustness). It can thus be inferred that the proposed framework can be employed across unlabelled datasets for classification of difficult to classify data points, therefore addressing one of the two key challenges in classification.

The next two chapters will address the second challenge in classification i.e. use external information to improve fusion step in ensemble classification. Comprehending the contribution of each classifier in ensemble can lead to robust (capability to adapt) and explainable (transparent) decision level fusion. The next chapter puts forward a new class of '*A priori*' FM and aggregation operator based ensemble classifier to tackle this challenge.

Chapter 5

Combination of Classifiers Through Leveraging External Knowledge

This chapter moves on from the previous two chapters to address a different category of challenge that affects ensemble classification. The chapter focuses on leveraging existing external information improving the overall classification performance as understanding the decision level fusion step in the ensemble could potentially lead to a robust (capability to adapt) and explainable (transparent) approach. In doing so, the chapter also addresses objective 3 of this thesis (for more details refer to section 1.2 in Chapter 1). Moreover, leveraging additional external information can be useful to understand how classifier outputs are actually combined. Currently, linear aggregation is state-of-the-art for combination of classifier outputs. However, these methods cannot account for non-linear relationships between the classifier outputs. Although, few non-linear aggregation methods have been explored [23, 94], these approaches do not leverage external information.

One approach to use the external knowledge is the aggregation operators (such as FIs), which are powerful non-linear mathematical functions (unlike most of the ensemble methods which perform linear combinations) capable of exploiting interactions between the models in the ensemble, through a FM. Current FMs suffer from some challenges or they cannot capture the external knowledge. Thus, in this thesis a new category of *A Priori* FM is introduced. Section 5.1 will discuss the background and motivation for capturing the additional external information and drawbacks of current FMs (together they highlight the need for a new category of FMs). Section 5.2 will introduce this new category of *A Priori* FM.

The Choquet Fuzzy Integral (CFI) has been one the most widely used aggregation operator in literature. However, it suffers from some drawbacks leading to non-intuitive results. Recently, Wagner et al. [7] discussed these problems in detail and introduced a

new aggregation operator called the Recursive Average (RAV), which could be useful to understand how these sources (individual classifiers in the case of RAV operator based ensemble classification) are combined. With the introduction of RAV, there is a need to examine the most appropriate aggregation operator (between CFI and RAV) for the *A Priori* FM and thus, Section 5.3 will compare CFI and RAV.

5.1 Background and Motivation

One important challenge to achieve better ensemble classification performance i.e. to better combine the individual classifiers output. To tackle the aforementioned challenge in ensemble classification, one approach is the use of the aggregation operator defined with respect to a Fuzzy Measure (FM) [159, 160, 23]. Aggregation operators are mathematical methods which have been extensively used in the literature for weighted multi-source information fusion [159, 160] (for more details refer to Section 2.4 in Chapter 2). The weights of the sources (e.g. individual classifier outputs in the context of ensemble classification) are defined by a Fuzzy Measure (FM) [23, 91], which not only captures the worth of the individual sources, but also the weights of all the possible combinations of sources (again, more details in Chapter 2).

To solve real-world problems using aggregation operators, the weighting strategy needs to be defined in advance. For example, defining FMs with respect to FIs is an intrinsic task. Given n information sources, 2^{n-2} factors need to be identified to define a FM. Therefore, defining a FM is a complex expensive task. Currently, to construct a FM, three major approaches are common: 1) Experts-driven, 2) Optimisation and 3) Algorithm-driven. These methods to construct FMs have been widely used in literature [83, 87, 23], however, each of these methods have their distinct strengths and weakness. With increase in number of sources, the number of factors to be defined for a FM increase exponentially; thus, manual specification of every combination sometimes becomes impractical. A body of existing literature has addressed FMs learned from training data with respect to a selected aggregation operator and an optimisation approach. However, these FMs are non-generic and should only be using the same aggregation operator defined in advanced. Algorithmic versions of FMs such as the Sugeno λ-measure and the S-decomposable measure, construct the entire FM lattice from densities (worth of the singletons) and thus, they fail to capture the interaction among the sources. Moreover, they are commonly driven by mathematically required properties, rather than application-specific insight.

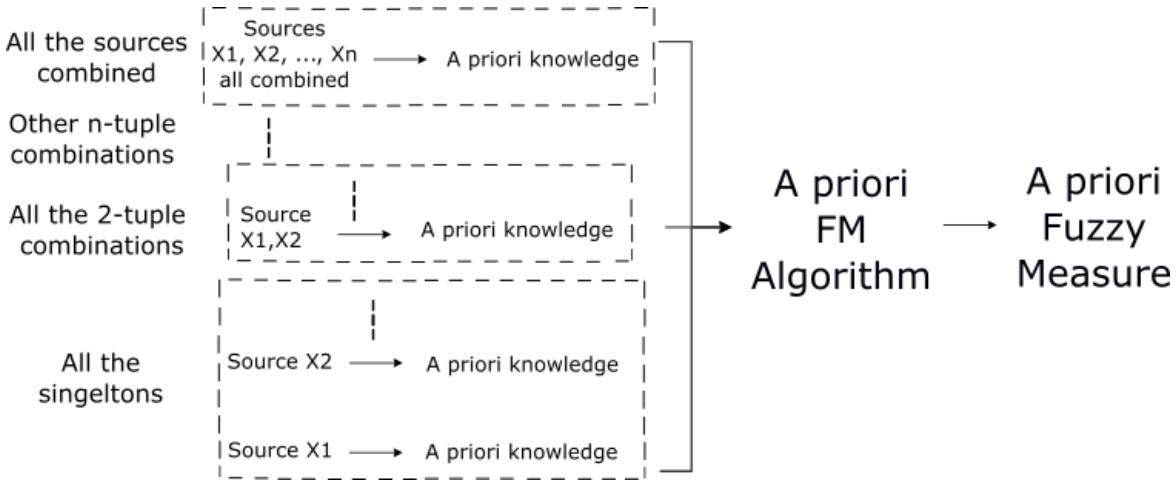
The work in this chapter addresses the aforementioned drawbacks and proposes a fourth category: FMs specified using externally available information, so-called '*A Priori* FM', an alternative method to construct the FM by using information from external available information (e.g. literature) i.e. from *'a priori'* knowledge. Here, 'external information' is information available to inform the weighting of individual sources and their combinations which is independent of the actual data fusion step (e.g. classifier combination in ensemble classification). The next section introduces this '*A Priori*' FM along with two approaches to capture the external information to drive the proposed FM.

5.2 *A Priori* Fuzzy Measure

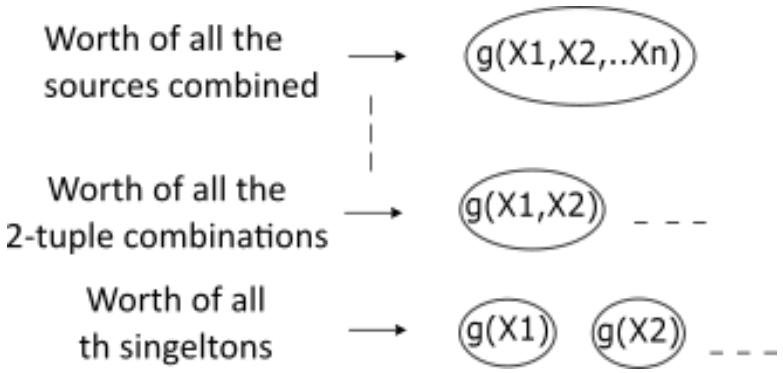
It is important to recall that FMs can capture different levels of knowledge and different knowledge content by evaluating interactions among different information sources. This section puts forward the novel concept of *A Priori* FMs - which are complex, yet human-interpretable weighting structures and parameterise aggregation functions such as the FI by leveraging existing external knowledge. The key idea underpinning this FM is to use external information for a robust (can adapt for different information sources) and interpretable (explainable) combination of sources. The A Priori FM also shares similar motivation with Fuzzy Logic Systems i.e. the idea of leveraging the external information from literature in order to construct intelligent systems. One could argue that expert-driven FMs are also a type of *A Priori* FM; nevertheless, as expert-driven FMs are a specific type with a long and well understood tradition/rationale, we argue that maintaining them as a separate category is both useful and serves clarity.

To achieve better overall ensemble classification results, this section employs this aforementioned novel *A Priori* FM to better combine individual classifier outputs. In particular, this thesis explores the decision-level aggregation of classifiers and focus on advanced fusion operators and their configuration as a part of this. In addition to leveraging existing external knowledge, this new *A Priori* FM based ensemble classifier can also enable interpretability in the fusion step of ensemble (based on the external information).

The next sub-section presents an algorithm to convert the external information to *A Priori* FM, followed by an *A Priori* FM based ensemble classifier - which can capture the quality of individual classifiers (and their combinations) in order to then enable a fusion-based (ensemble) classification.



(a) Steps of constructing FM from performance measures



(b) Lattice diagram showing the worth at each level

Figure 5.1 Proposed Fuzzy Measure Construction Framework.

5.2.1 Algorithm to convert external information to Fuzzy Measure

This section introduces a novel algorithm which converts the obtained information values to the *A Priori* FM, pictorially described in Figure 5.1a. For all $A \subseteq X$ and number of information sources $N: \{1,..,n\}$, the *A Priori* FM is given by g_{ap}. Algorithm 2 takes as input the information values $iVal_i$, where i is all the combinations, and the normalisation factor N_f (the range on which the accuracies are normalised); and generates the *A Priori* FM g_{ap} by running the following two steps:

S1: The first step is to map each information value to a normalised measure value *nm*, using step 'S1' shown in Algorithm 2. This step is done to apply the boundary property 'P1' of the FM (defined in Chapter 2 Section 2.4). The normalisation method is carefully chosen, as it maps the confidence in information for all the individual sources and all

the subsets of sources, against the overall maximum information. Two normalisation approaches are used in this work, described as follows:

1. One approach is to normalise the *iVals* between the minimum and the maximum observed information value (min-max normalisation) i.e. the $N_f = 0$. This approach is chosen as it is one of the most commonly used approach in literature.
2. Another approach is to normalise the *iVals* between 50% (of highest possible value which could be 1 or 100 depending upon application) and the maximum observed information value (*MaxiVal*) i.e. use the normalisation factor $N_f = 50$. The normalisation factor is chosen to be 50% (instead of 0) as intuitively the applications should perform better than random guessing i.e. their information value should be more than 50%.

These normalised measure values are then subtracted from one (to give high worth to best *iVal* and vice-versa) and then passed onto the second step.

S2: The second step is to convert the normalised measure values to the *A Priori* FM. For example, for all the individual sources from Figure 5.1b i.e. i.e.g. $g\{x_1\}$, $g\{x_2\}$, etc., the normalised measure values obtained from the previous step is the final *A Priori* FM. For each combination of sources (i.e. all the combinations in Fig 5.1b), if the normalised measure value obtained from the previous step is greater than or equal to the values of all its sub-sources, the normalised measure value of that combination is the final *A Priori* FM (i.e. from Fig 5.1b, if $nm(\{x_1,x_2\}) > nm(\{x_1\})$ and $nm(\{x_1,x_2\}) > nm(\{x_2\})$, then $g_{ap}(\{x_1,x_2\}) = nm(\{x_1,x_2\})$). Otherwise, if this normalised measure value is less than the measure value of any of its sub-sources, the maximum measure value among its sub-sources is the final *A Priori* FM (i.e. from Fig 5.1b, if $nm(\{x_1,x_2\}) < nm(\{x_1\})$ or $nm(\{x_1,x_2\}) < nm(\{x_2\})$ or both, then $g_{ap}(\{x_1,x_2\}) = max\,(nm(\{x_1\}),nm(\{x_2\}))$).

Information regarding an event (or system) can be collected from different sources. Although this knowledge can be collected in multiple ways, in this thesis we present two different approaches to capture the *a priori* knowledge as an external source of information to construct the FM (the entire lattice) and evaluate it in the context of ensemble classification. The two approaches are: 1) Classifiers performance driven, and 2) Literature knowledge driven. The first method constructs the FM from the training data of a contemporary empirical experiment result. By contrast, the second method constructs the FM from a pre-existing expert knowledge that is instead obtained from psychological experiments. A detailed

Algorithm 2: *A Priori* Fuzzy Measure Algorithm

inputs : Information Values ($iVal$), Normalisation factor (N_f)
output : Fuzzy Measure (g_{ap})

$MaxiVal = \arg\max(iVal)$
foreach i *in* $nm(A)$ **do**
  S1:
  $nm(A_i) = 1 - \frac{MaxiVal - iVal(i)}{MaxiVal - N_f}$
  S2:
  if $|A| == 1$ **then**
    $g_{ap}(A_i) = nm(A_i)$;
  else
    if $nm(A_i) < nm(B_i)$ **then**
      $g_{ap}(A_i) = nm(B_i)$;
    else
      $g_{ap}(A_i) = nm(A_i)$;
where, $nm(B_i) = max(g_{ap}(A - 1_i))$

explanation of the two approaches, followed by a decision level fusion ensemble classifier based on this novel *A Priori* FM are presented in the next three subsections respectively.

5.2.2 Classifier Performance Driven *A Priori* Fuzzy Measure Learning

In classifier performance driven FM learning, the *a priori* knowledge i.e. the information value *iVal*, is the classification performance of all the individual classifiers and the classification performance of all the combinations of the classifiers. This value can be metrics such as accuracy, specificity, sensitivity, AUC, etc. Using the Algorithm 5.2.1 introduced in last sub-section and the framework shown in Figure 5.1, the (observed) performance metric can be converted to the *A Priori* FM.

5.2.3 Literature Driven *A Priori* Fuzzy Measure Learning

In the aforementioned approach, the performance values (*a priori* knowledge) is obtained from empirical experiments. In contrast, the performance values in literature driven approach are obtained from literature, for example existing publications from other disciplines such as Psychology. These Literature driven FM are motivated by a pivotal question in aggregation operator based data fusion: 'Can we capture the FMs in cases where we don't have data to learn from (i.e. from algorithms or experts)'. The Literature-driven *A Priori* FM is based

on an experiment in literature, investigating how well the subjects (people) perform in the selected application. It is important to note that the framework (Figure 5.1) to construct the FM from performance values is the same in both the approaches and they differ in how the information is collected, emphasising that this method is a robust approach to construct the FM.

5.2.4 *A Priori* Fuzzy Measure Driven Ensemble Classifier

We now proceed to develop a class of ensemble classification algorithms, which use this novel *A Priori* FM to support the decision level fusion step in the ensemble. For simplicity, only binary classifiers are considered in this thesis. Let f_k be k^{th} binary classifier in the ensemble having n classifiers $(X_1, X_2, ..., X_n)$ and the two classes be labelled $y = +1$ and $y = -1$. The class labels $sign(f_k(x_i))$ obtained for each training data x_i is then used to compute the final decision value $h(x_i)$. The *A Priori* FM is computed using this decision value $h(x_i)$. Each test data x' can now be classified using the following steps:

1. Compute the decision value $h(x')$ for all the classifiers in the ensemble,
2. The aggregation value agg for the test data is computed using the *A Priori* FM g_{ap} in respect to the RAV (2.17) or the CFI (2.16).
3. The final class label is $sign(agg)$.

In the next section, we compare the most commonly used FI i.e. CFI with the recently introduced RAV with the intent of identifying the most appropriate aggregation operator for *A Priori* FM.

5.3 Aggregation operator selection

Many aggregation operators have been proposed in the literature, FIs [90] being the one of the most commonly used (especially the CFI). FIs are powerful non-linear aggregation functions and may result in better performance than the most commonly used weighted or non-weighted aggregation functions, such as the Simple Weighted Average (SWA) or Ordered Weighted Average (OWA). We focus on the FIs for fusion problems as unlike most of the aggregation methods which perform linear combinations, FIs are capable of capturing interactions between the information sources. FI based data fusion methods have found applications in numerous domains including Software defect prediction [97], Multi-Criteria

Decision Making (MCDM) [95], Brain Computer Interface (BCI) [98], Forensic science [96], and Explosive hazard detection [101]. One specific application of interest is Computer Vision [99, 135, 136, 100].

Although FIs have been used widely, they suffer from potential drawbacks of producing non-intuitive results in respect to a commonly intuitively interpretable FM. For example, as shown in [7], the re-ordering inherent to the FI can lead to only partial exploitation of a FM and the aggregation of symmetrically mirrored inputs does not necessarily result in symmetrically mirrored outputs. Wagner et al. recently discussed these aspects in detail in [7] and introduced a family of aggregation functions called the Recursive Average (RAV), as an alternative to FIs for fusion applications designed to leverage (interpretable) FMs. Moreover, due to the fusion principles of RAV, they can help in understanding how the sources are fused (in the case of ensemble classification application these sources are individual classifiers). This operator can therefore be useful to understand the combination step in ensemble classification.

With the RAV operator addressing some of the potential shortcomings of how FMs are used by FIs, there is renewed scope to revisit the most appropriate aggregation operator for the *A Priori* FM. Therefore, before presenting the case studies for this new FM across on ensemble classification applications (in order to understand the configurations of fusion step in ensemble classification), it is important to choose the appropriate aggregation operator. It is therefore intuitive to compare the two aggregation operators CFI and RAV, and in order to do so, the following two questions needs to be answered:

A Is there any difference in performance between the aggregation operators CFI and RAV?

B What do the resulting FMs for both the operators look like and how do they compare?

In order to answer these questions, an empirical comparison between the CFI and the RAV operators is vital. Subsequently, in the next section we will compare CFI and RAV operators based ensemble classifiers, which internally compares the two aggregation operators. Since RAV is relatively new, no RAV based ensemble classifier has been presented so far. In this chapter, we first present Decision-level Fusion of multiple kernels using RAV (DFRAV). DFRAV aggregates the outputs of kernel based classifiers through the use of the RAV with respect to a FM learned through a regularised quadratic programming approach. DFRAV is based on similar principles of DeFIMKL. The DeFIMKL ensemble classifier, which is a state-of-the-art CFI aggregation operator based ensemble classifier, has been reviewed in the Section 2.4.3 of literature review chapter. It will be therefore intuitive to compare these two

algorithms. It is interesting to note this comparison can inherently contextualises the two aggregation operators, potentially answering question A (defined above). The next section first puts forwards the DFRAV ensemble classifier, followed by its detailed comparison with the DeFIMKL.

5.3.1 Decision-level Fusion of multiple kernels using Recursive Average (DFRAV)

In this section, we propose Decision-level Fusion of multiple kernels using RAV [1] (DFRAV) and compare its performance with DeFIMKL, which is a state-of-the-art aggregation operator based ensemble classifier (for more details refer to section 2.4.3). Since FM defines the working of the RAV, thus the selection of FM selection is important for the working and understanding of the DFRAV algorithm. For the comparison to be consistent the FM for RAV is also constructed using a regularised sum of squared error (SSE).

The SSE for ensemble decision $RAV_{FM}(x_i)$ with respect to FM be defined as:

$$E^2 = \sum_{i=1}^{N} (RAV_{FM}(x_i) - y_i)^2, \tag{5.1}$$

where n = number of training samples, y_i is the actual class label for x_i.

Let $f_k(\mathbf{x_i})$ be the output generated by the kth classifier in an ensemble on a feature vector $\mathbf{x_i}$. The overall decision of the ensemble is computed by the RAV, where g encodes the relative worth of each classifier in the ensemble. Thus, the output of the ensemble with respect to the FM g on feature-vector $\mathbf{x_i}$ is produced by $RAV_{FM}(x_i)$, mathematically described as follows:

$$RAV_p(x_i) = \begin{cases} h(S), & \text{if } |S| = 1. \\ (\frac{\sum RAV_{FM}(B_j)}{c} . \sum g(B_j), & \text{otherwise.} \end{cases} \tag{5.2}$$

where $B_j = S \backslash x_j, x_j \in S$ and c = cardinality of B_j. The SSE can be expanded as:

$$E^2 = \sum_{i=1}^{N} (H_{x_i}^T . u - y_i)^2, \tag{5.3}$$

[1] For simplicity, we focus our experiments on p=1.

where y_i is the actual class label for x_i, u is lexicographically ordered FM g i.e. $u = (g\{x_1\}, g\{x_2\}, ..., g\{x_1 \cup x_2\}, g\{x_1 \cup x_3\}, ..., g\{x_1 \cup x_2 \cup ... \cup x_m\})$, and

$$H_{x_i} = \begin{pmatrix} h(S_1)/1 \\ h(S_2)/1 \\ . \\ . \\ (h(S_1)+h(S_2))/2 \\ . \\ . \\ (h(S_1)+h(S_2)+..+h(S_n))/n \end{pmatrix} \tag{5.4}$$

where H_{x_i} is of size $(2^m-1)X1$ and contains all the difference terms $f_{\pi(k)}(\mathbf{x_i}) - f_{\pi(k+1)}(\mathbf{x_i})$ at the corresponding locations of A_k in u. We can fold out the square terms from equation 5.3, producing:

$$\begin{aligned} E^2 &= \sum_{i=1}^{n} (u^T H_{x_i} H_{x_i}^T u - 2y_i H_{x_i}^T u + y_i^2) \\ &= (u^T D u + f^T u + \sum_{i=1}^{n} 2y_i H_{x_i}), \end{aligned} \tag{5.5}$$

$$D = \sum_{i=1}^{n} H_{x_i} H_{x_i}^T, \quad f = -\sum_{i=1}^{n} 2y_i H_{x_i}$$

Equation 5.5 is a quadratic function and thus we can add the constraints on u such that it represents a FM, producing a constraint QP. We can add the monotonicity constraint on u according to the properties P1 and P2 as $Cu \leq 0$, such that C is a vector representing monotonicity constraint $g\{x_1\} - g\{x_1 \cup x_2\} \leq 0$ (see [91] for more details on C). Thus, the full QP to learn FM u is:

$$\min_u \ 0.5u^T \hat{D} u + f^T u, \quad Cu \leq 0, \quad (0,1)^T \leq u \leq 1 \tag{5.6}$$

where $\hat{D} = 2D$. We test the performance using l_1 regularisation i.e.

$$\min_u \ 0.5u^T \hat{D} u + f^T u + \lambda ||u||_1, \tag{5.7}$$

The QPs at 5.6 provide a method to learn the FM u (i.e. g) from the training data. A new feature vector x', from a test set, can be classified as:

1 Compute the SVM decision value $f_k(x')$,

2 Apply the CFI at equation 2.16 with respect to the learned FM g,

3 Compute the class label using $sign(f_k(x'))$.

Equation 5.3 can be transformed to a constrained Quadratic programming (QP) problem by adding Monotonicity and Boundary Condition constraints on u. The output of the QP is the FM g. For test data, the decision value is computed by applying the RAV operator i.e. equation 2.17, with respect to the learned FM g described as below:

1 Compute the SVM decision value,

2 Apply RAV w.r.t. the learned FM g,

3 Compute the class label.

5.3.2 Results and Discussion

The following table compares DeFIMKL with DFRAV. These algorithms are also compared with $GAMKL_p$, a feature level kernel fusion algorithms which uses p-norm genetic algorithm approach for kernel ensembles i.e. a variation of FIGA (more details in [23]). The DeFIMKL and DFRAV is contextualised with GAMKL to demonstrate the difference in performance of feature level fusion ensemble classifiers (GAMKL) and decision level fusion ensemble classifiers (DeFIMKL and DFRAV). The table shows the accuracies of all the algorithms along with their standard deviations over 100 trials, for six UCI datasets. Five base kernels with $\lambda = 0.5$ and $\sigma = 0.5 - 1.5 \ / \ (number\ of\ features)$ for L_1 and L_2 regularisation on SVM kernels are selected. Maximum of $GAMKL-L_1$ and $GAMKL-L_2$ i.e. L_1 and L_2 regularised GAMKL, was compared with DeFIMKL and DFRAV. The Genetic Algorithms for GAMKL is run for 25 generations with a population of 31 chromosomes with 5-fold cross validation (more details in [23]).

$DeFIMKL-L_1$: DeFIMKL algorithm using L1 regularisation $DeFIMKL-L_2$: DeFIMKL algorithm using L2 regularisation $DFRAV-L_1$: DFRAV algorithm using L1 regularisation $DFRAV-L_2$: DFRAV algorithm using L2 regularisation $arg\ max\ \{GAMKL-L_1,\ GAMKL-L_2\}$: best accuracy among GAMKL-L1 and GAMKL-L2

The classification accuracies of the DFRAV and DeFIMKL algorithm are shown in Table 5.1. A two-sample t-test at 5% significance is used to determine the best algorithm(s) for each dataset. The t-tests compare the accuracy distribution of each algorithm with the

Table 5.1 Comparison statistics of DeFIMKL and DFRAV. For clarity, the best classifiers are bold and the classifiers performances which are statistically significantly indifferent than the best are in *italics*

Dataset	$DeFIMKL-L_1$	$DeFIMKL-L_1$	$DFRAV-L_1$	$DFRAV-L_1$	$max\{GAMKL-L_1, GAMKL-L_2\}$
Sonar	83.83 (5.61)	83.88 (5.4)	84 (4.85)	83.88 (5.48)	**86.9 (1.68)**
Dermatology	*97.41 (1.91)*	*97.45 (1.55)*	*97.38 (1.76)*	*97.49 (1.74)*	**97.9 (1.68)**
Wine	*99.7 (0.91)*	*99.7 (0.84)*	*99.7 (0.87)*	*99.7 (0.91)*	**99.7 (0.87)**
Ionosphere	94.72 (2.96)	95 (2.37)	94.7 (2.56)	94.89(2.58)	**96.4 (1.01)**
Ecoli	*97.34 (1.63)*	*97.69 (1.62)*	*97.4 (1.88)*	*97.46 (1.67)*	**97.76 (3.17)**
Glass	*94 (3.34)*	*94.07 (3.23)*	*94.16 (3.44)*	*94.07 (2.83)*	**94.19 (4.93)**

'MaxGAMKL-L1, GAMKL-L2'. Thus, more than one algorithm can be considered as best, shown in bold font for each dataset. The results indicate that DFRAV and DeFIMKL produced very similar results, which answers *question A* (defined in Section 5.3) i.e. comparison of the performance of the aggregation functions CFI and RAV.

To answer *question B* (defined in Section 5.3), the resulting FMs are shown in Figure 5.2. The figure shows the FM values for CFI represented by G_CFI, obtained after optimisation of equation 4 using quadratic programming. The figure also shows the variations of FM for RAV represented by G_RAV, obtained after optimisation of equation 7 using the quadratic programming. Let sources $\{1\}$, $\{2\}$, $\{3\}$, and $\{4\}$ be called FMs at level 1, similarly let $\{1,2\},\{1,3\}..\{3,4\}$ be FMs at level 2, and so on. Three sources in level 1 are very small for CFI while the fourth one is much larger. Although, the order of values is different, the same pattern is followed for RAV i.e. three sources are very small for RAV, and the fourth one is much larger. At level 2, three FMs are very small while the other three are relatively larger for both the operators. A similar trend is followed by the other two levels also. Although, the FMs are not in the same order, the trend of variations in values is similar at different levels.

5.4 Summary

To generate a robust and better '*ensemble*' classification method, we propose to leverage additional external information about the data. The use of this additional knowledge can also help to understand the fusion step in ensemble classification. One approach to capture this information is the use of FI aggregation operator w.r.t. to a FM. Three types of FMs exist in the literature, however, due to their methodology they either could not be used or cannot capture this external information. With an increase in lattice size, manually defining FMs becomes virtually impossible. Algorithm based methods do not capture interactions

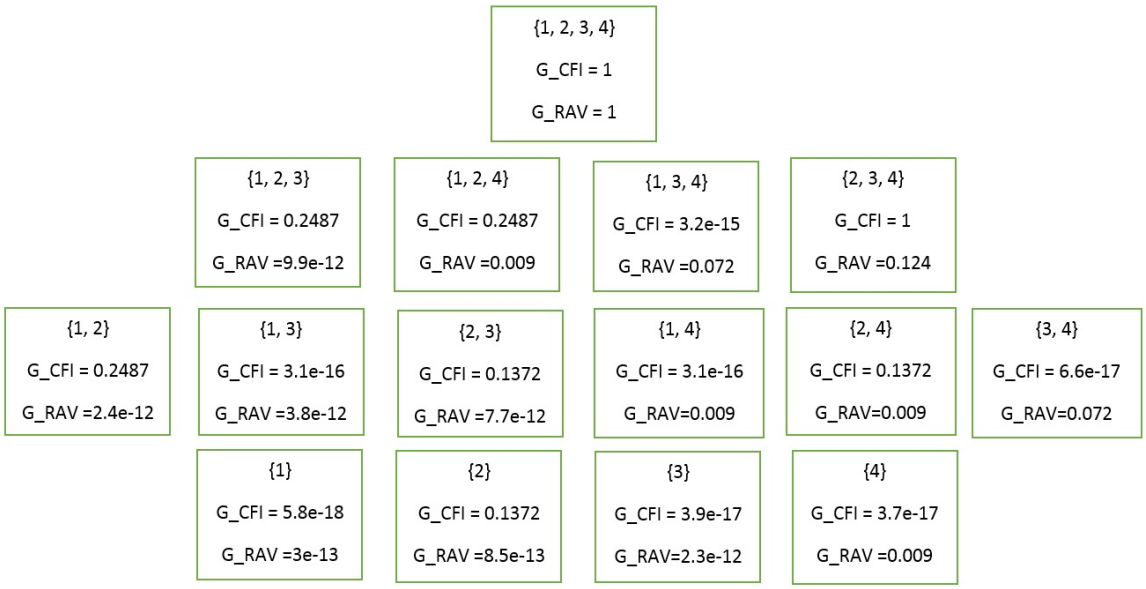


Figure 5.2 FM values comparison of CFI and RAV

among the individual sources as they use some form of mathematical computation on these individual sources. Lastly, optimisation based methods cannot be used to capture external information. Therefore, this chapter introduced a fourth category of FMs, the *A Priori* FM, to capture the external information.

Another important factor is to select the most appropriate aggregation operator for the *A Priori* FM. FIs are most commonly used aggregation operator, although they suffer from potential drawbacks detailed in [7], resulting in non-intuitive outcomes when employed in respect to interpretable FMs, i.e. FMs which can be interpreted to provide meaningful insight into the value of information sources and their combinations. The RAV aggregation operator was introduced as an alternative to FIs, specifically for cases where a FM is available which is independent from the aggregation operator, for example a FM which is specified using external information (cf. experts). To select the most appropriate aggregation operator, it is imperative to compare the CFI with the RAV. Subsequently, a decision-level fusion algorithm called DFRAV was proposed, which combined this new FM using RAV aggregation operator. DeFIMKL was then compared with DFRAV, which principally compares the two aggregation functions CFI and RAV. The results from Table 5.1 show that RAV and CFI achieve equivalent performances on six standard datasets, indicating that both the aggregation operators could be used with the proposed *A Priori* FM.

Having introduced the *A Priori* FM, it is important to evaluate how this FM performs in different applicaitons. The next chapter will therefore evaluate this *A Priori* FM, for both the CFI and the RAV aggregation operators, across three empirical case studies.

Chapter 6

A Priori FM driven ensemble classifiers

To capture the additional external information present in the data, the *A Priori* FM has been introduced in the last chapter. Three case studies are employed which empirically examine the characteristic details of the decision-level fusion step in the ensemble classification (addressing objective 5 stated in Section 1.2 of Chapter 1). To achieve this, the chapter focuses on a systematic evaluation and demonstration (across different applications) of an *A Priori* FM based ensemble classifier, using both the CFI and the RAV aggregation operators. Section 6.1 presents the first case study within a face recognition application i.e. to recognise subjects by combining multiple individual classifiers trained on parts of their faces. The second case study in Section 6.2 presents a comprehensive study comparing the *A Priori* FM with Uriz FM and Sugeno λ-measure for both the CFI and RAV aggregation operator, across 20 datasets from the UCI repository. Lastly for the third case study in Section 6.3, the *A Priori* FM based ensemble classifier is contextualised to a wider literature by comparing it with state-of-the-art aggregation operator-FM based ensemble classifier DeFIMKL and traditional ensemble classifiers such as Adaboost, Bagging, Random Forest and Majority Voting (for 20 datasets).

6.1 Case Study 1: Generating *A Priori* Fuzzy Measures from Knowledge for Face Recognition

6.1.1 Problem Overview

In computer vision related literature, various methods [130, 137] exist for multi-region face recognition application i.e. for recognising faces by combining classifiers which individually

focus only on a specific part of the face. A number of researchers have used different versions of Convolutional Neural Networks (CNNs) to tackle this problem. Recently, the Multi-Region Ensemble Convolutional Neural Network (MRE-CNN) method has been a powerful approach in respect to similar problems such as age estimation [137], when local face features are combined with the whole face.

To construct the A Priori FM, the multi-region facial recognition is employed as a case study i.e. we focus on systematically combining the individual classifiers which are trained on parts of the face. To achieve this, we use a decision-level fusion approach which employs the A Priori FM and configure it with a priori information. Two ways of building the A Priori FM are explored and evaluated: 1) Classifier performance driven learning, and 2) Literature driven (for more details refer to Section 5.2 of Chapter 5).

For the classifier performance driven approach, the classification results of the individual face regions and the classification results of all the combinations of the regions are used to construct the A Priori FM. By contrast, for literature driven A Priori FM, the a priori information is obtained from the expert knowledge e.g. psychological experiments (more details in the next section). As such, the focus of this case study is not to examine which a priori information or which approach works best, but to explore and demonstrate how well aggregation operator based classification approach can perform when its FM is given additional *a priori* information.

6.1.2 Experiment

To construct the *A Priori* FM, the application of multi-region facial recognition (i.e. to fuse multiple face regions towards one superior classifier) is used. Different face features have different levels of information and different information content. Moreover, this information can be collected from both the empirical evaluation and the expert knowledge. Therefore, this application makes an ideal case study to demonstrate both the classifier performance and literature driven A Priori FM approach, for the CFI and RAV aggregation operators.

To perform the experiments, the AT&T dataset [11] containing a total of 400 images for 40 subjects (ten different images for each subject) with various illumination, pose and expressions, was selected. The faces were cropped into three different regions: Eyes (including eyebrows), Nose (including cheeks) and Mouth (shown in Figure 6.1). These regions are selected as they are the most important for face recognition and reflect real world settings. Employing the multi-region facial recognition application on AT&T dataset as the knowledge sources, the experimental settings to construct the *A Priori* FM for both the



(a) Sample original face image (b) Cropped face image

Figure 6.1 Sample image from AT&T dataset (replicated from [11])

approaches (the classifier driven and the literature driven) are explained in the following two sections.

6.1.3 Classifier Performance Driven *A Priori* Fuzzy Measure Learning: CNN experiment based accuracy

The first approach to construct the A Priori FM is driven by classifier performance values. Since Convolutional Neural Networks (CNNs) [131, 132] are state-of-the-art methods for classification across face recognition applications [138, 137] (for more details refer to Section 2.5.3 Chapter 2), they are used as the base classifiers in this case study. To construct the classifier performance driven *A Priori* FM, three CNNs trained on each atomic facial feature of all the training set images from the AT&T dataset are chosen as the external knowledge source. The accuracies of all the combinations of these facial features are computed by combining the accuracy of the individual CNNs trained on the respective features e.g. the accuracy for combination of the eyes and the nose is computed by adding the probabilities of the two CNNs trained on these two features (for clearer understanding, a sample computation is shown at the end of this section). Lastly, the *A Priori FM* Algorithm (Algorithm 2 from Chapter 5) is run to convert these observed accuracies to the final FM.

Table 6.1 Observed *A Priori* performance from the two experiments

Feature	Accuracy on AT&T database (in %age)	Accuracy from human recognition (in %age)
Eye	83.08	91.6
Nose	76.42	71.7
Mouth	75.67	75.3
Eye + Nose	91.25	93.3
Eye + Mouth	90.83	95.5
Nose + Mouth	86.08	78.3
Eye + Nose + Mouth	92.58	93.3

The three CNNs employed are all three layered, having two convolutional layers and one final fully connected layer. The networks are trained using a stochastic Adam optimisation method [161] with a learning rate of 0.009 and 100 training epochs. The three classifiers give the performance for each atomic face region and subsequently the performance of all the combinations are computed by merging the probabilities of three base classifiers. For each trial, 70% of the data from AT&T dataset were randomly used for training the classifiers and the remaining 30% for testing. In these 30 trials, we have made sure that all data has been selected for training in several trials as well as tested several times. The second column in Table 6.1 shows the observed training accuracy, which is used as the external knowledge for the *A Priori* FM.

Upon obtaining these empirical accuracies, two normalisation approaches (discussed in detail in Section 5.2 of Chapter 5) are used to construct the *A Priori* FM for both the CFI and the RAV aggregation operators. The first approach for is to use normalisation factor $N_f = 50$, represented as g_{ap1} in the Figure 6.2 and $aPriroi_{ap1}$ in Table 6.2 (introduced in Section 5.2.1 in Chapter 5). The second approach is to replace 50% by 0 i.e. $N_f = 0$, represented by g_{ap2} in Figure 6.2 and $aPriori_{ap2}$ in Table 6.2 (introduced in Section 5.2.1 in Chapter 5). The final classifier performance based *A Priori* FM is shown in Figure 6.2, and the accuracy on the test set is shown in the Table 6.2. It is imperative to contextualise the performance of *A Priori* FM by comparing it with other established methods form the literature. Thus, in addition to the ensemble classifier driven by *A Priori* FM, the table also shows the classification results for Uriz FM and the Sugeno λ-measure driven ensemble classifiers. Uriz FM is selected as it is has similar motivations to *A Priori* FM as they both use performance values (Uriz FM is limited to classifier accuracy values while *A Priori* FM is a generic) to construct the FM, however they are distinctly different in how they construct the FM. The *A Priori* FM is also

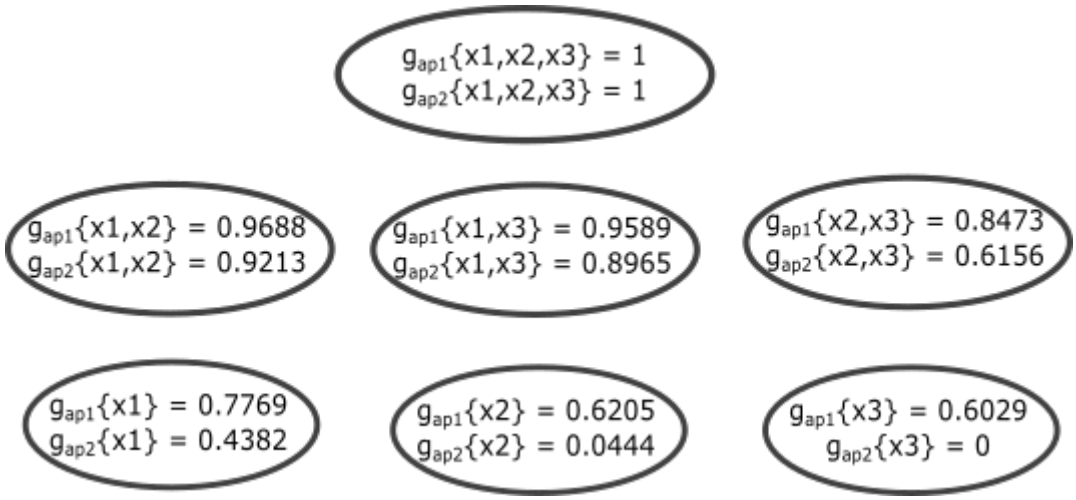


Figure 6.2 Fuzzy Measure Lattice for classifier performance driven accuracy

compared with Sugeno λ-measure as it is one of the most commonly used method in the literature.

The Table 6.2 shows the performance of the two *A Priori* FMs, the Uriz FM and the Sugeno λ-measure based ensemble classifiers for two aggregation operators, the CFI in second column and the RAV in third column respectively. It can be observed that the two *A Priori* FMs and Uriz FM based ensemble classifiers outperform the Sugeno λ-measure based ensemble classifiers, for both the aggregation operators. With CFI the *A Priori* FM with $N_f = 50$ i.e. $aPriroi_{ap1}$ gives best result, whereas for RAV the best performance is obtained for the Uriz FM based ensemble classifier. The average and weighted average classification accuracy of the three atomic features is 81.67% and 81.85% respectively, inferior to the combination of the three features reported in the table. It can be thus inferred that the *A Priori* FM based ensemble classifiers results in good performance for both the aggregation operators.

Another approach to towards the external information is Literature-driven, discussed in the next section. Before presenting the approach, the *A Priori* FM computation using Algorithm 2 for a sample set of information values for three sources is shown below:

The A Priori FM generation using a sample set of information values (*iVal*) is presented below:

$$iVal_{x1} = 76.7$$

$$iVal_{x2} = 70$$

$$iVal_{x3} = 78.33$$

Table 6.2 Accuracy (in %age) comparison of the classifier performance driven A Priori FM with Uriz and Sugeno λ-measure for both the CFI and the RAV

Method	With CFI operator	With RAV operator
$aPriroi_{ap1}$	94.17	93.33
$aPriroi_{ap2}$	93.33	94.17
Uriz method FM	93.33	95
Sugeno λ FM	92.50	92.50

$$iVal_{x1,x2} = 85.83$$

$$iVal_{x1,x3} = 88.33$$

$$iVal_{x2,x3} = 77.5$$

$$iVal_{x1,x2,x3} = 87.5$$

Upon getting the information values, the next step is is to run the A Priori FM Algorithm (Algorithm 2 from Chapter 5) on these values. For demonstration, $N_f = 50$ is chosen for step 'S1'.

$$nm_{A_i} = 1 - \frac{MaxiVal - Per(i)}{MaxiVal - 50}$$

maximum of accuracies ($MaxiVal$) = 88.33

The normalised information values obtained after running this step are shown below:

$$nm_{x1} = 0.6966$$

$$nm_{x2} = 0.5218$$

$$nm_{x3} = 0.7391$$

$$nm_{x1,x2} = 0.9348$$

$$nm_{x1,x3} = 1$$

$$nm_{x2,x3} = 0.7175$$

$$nm_{x1,x2,x3} = 0.9783$$

Step 'S2' of the algorithm is to convert the normalised information values to the FM. For base sources (densities), the FM is equal to the normalised information value, as shown below:

$g_{x1} = 0.6966$

$g_{x2} = 0.5218$

$g_{x3} = 0.7391$

For the remaining sources (combinations), if the normalised information value is equal or greater than its sub-sources (i.e. or individual source or combinations of sources from level below), the normalised information value is the FM. However, if the information value is less than its sub-sources, assign the maximum normalised information value from its sub-sources as the FM for that (e.g. $nm_{x2,x3} = 0.7175 \leq$ either of $(nm_{x2} = 0.5218, nm_{x3} = 0.7391)$, therefore, $g_{x2,x3} = nm_{x3}(0.7391)$). The remaining values of the A Priori FM lattice are shown below:

$g_{x1,x2} = 0.9348$

$g_{x1,x3} = 1$

$g_{x2,x3} = 0.7391$

$g_{x1,x2,x3} = 1$

6.1.4 Literature Driven *A Priori* Fuzzy Measure Learning: Psychological experiments based information values

The second external information source to construct the Literature-driven A Priori FM by using a priori knowledge from human face recognition (psychological) experiment. In this case study, the *a priori* knowledge is based on experiment in psychology, investigating how well the subjects (people) classify faces when seeing part of the face. The focus is to construct the FM from expert knowledge of psychological experiments and drive this knowledge for intelligent information fusion. It is important to reiterate a very important difference from the previous method i.e. in this approach the FM is constructed from information obtained from the pre-existing expert knowledge while in the aforementioned approach (classifier performance driven) the FM is constructed from the empirical results.

The information values for this experiment are obtained from the study by Matthews [162]. In the selected experiment, ten subjects were shown 200 images of feature changes for each atomic region and the combination of regions. The third column in Table 6.1 shows the mean information values (accuracy in this case study) of these ten subjects for each feature combination.

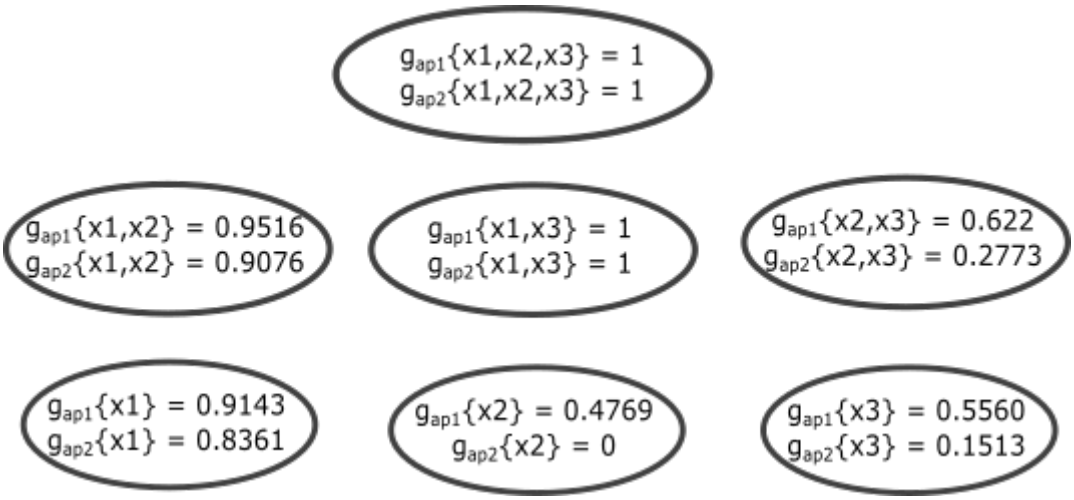


Figure 6.3 Fuzzy Measure Lattice for literature knowledge driven accuracy

The A Priori FM Algorithm is now used to convert the accuracies from psychological experiment to the A Priori FM. Similar to classifier performance driven A Priori FM, the two same normalisation approaches i.e. $N_f = 50$ and $N_f = 0$ are used. g_{ap1} and g_{ap2} in Figure 6.3 shows the A Priori FM values using normalisation values $N_f = 50$ and $N_f = 0$ respectively.

Table 6.3 compares the final classification accuracy of the A Priori FM based ensemble classifier with the Uriz FM and the Sugeno λ-measure based ensemble classifiers, for both the CFI and the RAV aggregation operators. $aPriroi_{ap1}$ and $aPriroi_{ap2}$ represent the final prediction for *A Priori* FM using normalisation values $N_f = 50$ and $N_f = 0$ respectively. Uriz method FM and the Sugeno λ FM represent the final result obtained using the Uriz FM and the Sugeno λ-measure respectively.

From the Table 6.3 it can be observed that the *A Priori* FM with $N_f = 0$ and the Uriz FM based ensemble classifiers outperform the *A Priori* FM with $N_f = 50$ FM and the Sugeno λ-measure based ensemble classifiers, for both the CFI aggregation operator. For the RAV aggregation operator, the Uriz FM based ensemble classifier outperforms all the other three methods. Specifically comparing the two *A Priori* FMs for both the aggregation operators, the $aPriroi_{ap2}$ slightly performs better than the $aPriroi_{ap1}$. Comparing the aggregation operators, overall the RAV performs better than the CFI. A more detailed discussion on the results for the decision level fusion using the literature driven and classifier performance driven A Priori FM based is discussed in the next section.

Table 6.3 Accuracy (in %age) comparison of the literature driven A Priori FM with Uriz and Sugeno λ-measure for both the CFI and the RAV

Method	With CFI operator	With RAV operator
$aPriroi_{ap1}$	93.33	94.17
$aPriroi_{ap2}$	94.17	94.17
Uriz method FM	94.17	95
Sugeno λ FM	92.50	93.33

6.1.5 Discussions

This case study evaluates the proposed *A Priori* FM based ensemble classifier within a multi-region face recognition application. Two approaches i.e. $aPriroi_{ap1}$ and $aPriroi_{ap2}$ convert the information values to the A Priori FMs. To contextualise the performance of the A Priori FM, we compare them to two well established measures i.e. the Uriz FM and the Sugeno λ-measure. From the results it can be observed that the A Priori FM and the Uriz FM outperform the state-of-the-art Sugeno λ-measure for both classifier driven and literature driven methods and for both the CFI and the RAV aggregation operators.

For classifier driven approach, the $aPriroi_{ap1}$ outperforms every other FMs with CFI aggregation operator, whereas the Uriz FM outperforms every other FMs with RAV aggregation operator. For literature driven approach, the $aPriroi_{ap2}$ and Uriz FM have the best performance with the CFI aggregation operator while the Uriz FM outperforms with the RAV aggregation operator. From the two set of results it can be concluded that the A Priori FM and Uriz FM outperform the Sugeno λ-measure, and have the potential to become state-of-the-art approach to construct the FM. More case studies need to be performed to affirm this claim, performed in the later sections i.e. case studies 2 and 3.

6.1.6 Summary

With the aim to leverage additional external knowledge, in the last chapter a new approach to capture the FMs i.e. the *A Priori* FM, was introduced. The chapter also presented two different ways to capture the external knowledge in order to drive a decision level A Priori FM based ensemble classification. In this section we present a case study to evaluate the proposed A Priori FM based ensemble classifier within a multi-region face recognition application. Two different ways to capture the external knowledge i.e. classifier performance from CNNs and literature-driven from psychological experiments, are used in this case study.

The $aPriroi_{ap1}$ and $aPriroi_{ap2}$ *A Priori* FMs are compared with the Uriz FM and the Sugeno λ-measure for the At&T face recognition dataset. It was observed that the A Priori FM and Uriz FM performs best for the CFI aggregation operator while Uriz FM performs best for the RAV operator. Overall, both the A Priori FM and the Uriz FM perform better that the state-of-the-art Sugeno λ-measure. Moreover, from Figures 6.2 and 6.3 the importance classifiers in the final output can be comprehended, leading to interpretable results. From the two experiments it can be inferred that the use of external knowledge can perform well with capability to drive intelligent information fusion and lead to interpretable results. To further evaluate the *A Priori* FM across more applications, the next section presents its application on multiple datasets from the UCI repository.

6.2 Case Study 2: *A Priori* Fuzzy Measures for multiple UCI datasets

6.2.1 Problem Overview

In order to further evaluate the proposed *A Priori* FM, an in-depth experiment is conducted on 20 datasets form the UCI repository. Similar to last case study, we will contrast the proposed A Priori FM with both the Sugeno and Uriz FMs, for both the CFI and the RAV aggregation operators. Before presenting the experiments, a brief overview of the 20 UCI datasets is explained in the next sub-section.

6.2.2 Datasets and Pre-processing

20 benchmark datasets from the UCI machine learning repository [150] were selected to compare the performance of the selected algorithms, as shown in Table 6.4. These datasets are selected as they contain a different range of instances with a different type of datasets. Not all the UCI datasets are binary and thus in some cases, multiple classes are joined together for the purpose of binary classification [45].

The next step is data pre-processing to standardise the datasets for an unbiased comparison. All the data which had missing values were deleted and made homogeneous, i.e. all numeric. This might also help to locate inconsistencies among the data. Each dataset was processed using z-score [142] normalisation i.e. zero mean and unit standard-deviation. No further processing techniques were used as: 1) the aim of this work is not to report the best possible result for each dataset, but to compare the performance of the classifiers. 2) to improve the

Table 6.4 Dataset Description

Dataset	Binary Classes	Number of Features	Number of Instances
Dermatology	{1,2,3} vs {5,6,7}	33	366
Wine	{1} vs {2,3}	13	178
Ecoli	{1,2,5,8} vs {3,4,6,7}	7	336
Glass	{1,2,3} vs {5,6,7}	9	214
Sonar	{1} vs {2}	60	208
Ionosphere	{0} vs {1}	34	351
SPECTF Heart	{0} vs {1}	44	267
Bupa	{1} vs {2}	6	345
WDBC	{M} vs {B}	30	569
Haberman	{+} vs {-}	3	306
Pima	{+} vs {-}	8	768
Australian	{0} vs {1}	14	690
SA Heart	{0} vs {1}	9	462
Satimage	{1,2,3} vs {4,5,6,7}	36	6,435
Segmentation	{1,2,3,4} vs {5,6,7}	19	2,310
Mammographic	{0} vs {1}	5	830
Credit-approval	{+} vs {-}	15	653
Ozone	{0} vs {1}	72	1,848
Tic-tac-toe	{+} vs {-}	9	958
Ilpd	{1} vs {2}	7	583

classification results, further processing specific to each dataset would be required, leading to a more challenging comparison [45].

6.2.3 Results and Discussion

In this section, the results and discussion are presented for the comparison of the *A Priori* FM with Uriz and Sugeno FMs. Table 6.5 and Table 6.6 present the results of the *A Priori*, Uriz and Sugeno FM based ensemble classifiers for both the CFI and the RAV aggregation operators on the 20 UCI datasets. The first two columns report the performance of the Sugeno λ-measure based ensemble classification algorithm for both the CFI and the RAV aggregation operators respectively. Similarly, the third and fourth column capture the results for the Uriz FM based ensemble classification algorithms and the fifth and the sixth columns for the *A Priori* FM based ensemble classifiers. It is important to note that these classification accuracies are reported by taking an average over 100 runs, along with their standard-deviation in the brackets. During each run, 80% of the data is randomly selected for

training i.e. constructing the FM, and the remaining 20% for testing. For every dataset, the accuracy of each algorithm was compared with the accuracy of the best algorithm (shown in bold) using a two-sample t-test (at $p<.05$). Results of classifiers which are statistically not different to the best algorithm at captures in italics and the classifiers which are statistically significantly worse than the best are underlined.

The ensembles in Table 6.5 use the same base classifiers, i.e. five SVMs, whereas the ones reported in Table 6.6 use five different base classifiers, i.e. SVM, Decision Tree, Adaboost, Bagging and Neural Network. The *A Priori* FM based aggregation operators (both CFI and RAV) performed best across the five SVM base classifiers and thus are compared against state-of-the-art ensemble classification algorithms: DeFIMKL, MJSVM, Adaboost with trees, Bagging and Random Forest, as reported in Table 6.7.

From Table 6.5 it can be observed that nearly all classifiers result in performance which is not statistically different from the best classifier (hence nearly all results are shown in *italics*). Note, we refer to the *best* performance of a classifier when said classifier produces results which are either the best or statistically not different from the best.

The CFI based ensemble classifiers performed very similarly for all three measures, being *best* in 16 out of 20 datasets. The RAV operator based ensemble algorithm also had a very similar performance, being *best* among 15 datasets for the Sugeno and the *A Priori* FM, and 16 datasets for the Uriz measure. Overall, these algorithms outperformed all ensemble algorithms based on different base classifiers shown in Table 6.6. Here, CFI based ensemble classifiers were *best* in 2, 14 and 2 datasets for Sugeno, Uriz and *A Priori* measure respectively. Conversely, RAV based ensemble algorithms were *best* in 16, 15 and 17 datasets for Sugeno, Uriz and *A Priori* measures respectively.

6.2.4 Summary

In this case study we present an in-depth evaluation of the *A Priori* FM by comparing it with the well established Sugeno λ-measure and the Uriz FM; for both the CFI and the RAV aggregation operators. A decision level fusion based ensemble classifier driven by these measures is compared for 20 datasets from the UCI repository. The ensemble classification algorithms are constructed for two sets of base classifiers: a set of five SVMs classifiers and a mixed set of SVM, Decision Tree, Adaboost, Bagging and Neural Network classifiers. The three FMs – integrated with the different aggregation operators – were compared for 20 datasets from the UCI repository. Further, the *A Priori* FM ensemble algorithm with same

Table 6.5 Summary performance statistics of the *A Priori* FM, the Uriz FM and the Sugeno λ-measure based ensemble classifiers for the same base classifiers. For clarity, the best classifiers are bold, the classifiers performances which are statistically significantly indifferent than the best are in *italics* and the ones which are statistically significantly worse than the best are underlined.

Datasets	Sugeno CFI	Sugeno RAV	Uriz CFI	Uriz RAV	*A Priori* CFI	*A Priori* RAV
Dermatology	96.85 (2.07)	*97.14 (1.98)*	*97.16 (1.94)*	**97.18 (1.94)**	96.81 (2.16)	*97.15 (1.97)*
Wine	98.86 (2)	*99.16 (1.66)*	*99.16 (1.66)*	*99.16 (1.66)*	98.86 (2)	**99.16 (1.66)**
Ecoli	*96.87 (2.08)*	*96.71 (2.13)*	*96.68 (2.13)*	*96.68 (2.13)*	**96.88 (2.1)**	*96.71 (2.13)*
Glass	*92.91 (3.95)*	92.43 (4.32)	92.41 (4.3)	92.34 (4.34)	**92.93 (3.95)**	92.43 (4.32)
Sonar	*80.42 (6.1)*	*80.35 (6.33)*	*80.53 (6.4)*	**80.56 (6.21)**	*80.51 (5.99)*	*80.44 (6.34)*
Ionosphere	*94.68 (2.65)*	94.1 (2.59)	94.13 (2.63)	94.13 (2.63)	**94.69 (2.66)**	94.13 (2.61)
SPECTF Heart	**80.13 (5.23)**	*80.09 (5.2)*	*80.02 (5.09)*	*80 (5.08)*	*80.11 (5.19)*	*80.06 (5.19)*
Bupa	*67.35 (5.16)*	*67.74 (5.39)*	*67.75 (5.37)*	*67.65 (5.27)*	*67.36 (5.23)*	**67.75 (5.34)**
WDBC	*96.72 (1.79)*	*96.83 (1.79)*	*96.83 (1.78)*	*96.83 (1.83)*	*96.7 (1.79)*	**96.83 (1.78)**
Haberman	*73.03 (4.7)*	*73.19 (4.8)*	*73.19 (4.8)*	*73.21 (4.83)*	*73.05 (4.68)*	**73.21 (4.81)**
Pima	*75.94 (2.86)*	*75.95 (2.82)*	*75.95 (2.82)*	**75.97 (2.84)**	*75.95 (2.99)*	*75.95 (2.83)*
Australian	*84.86 (3.03)*	*85.13 (2.93)*	*85.14 (2.96)*	*85.13 (2.92)*	84.75 (3.01)	**85.14 (2.94)**
SA Heart	*70.99 (4.11)*	*71.13 (4.13)*	*71.13 (4.11)*	**71.18 (4.13)**	*70.95 (4.08)*	*71.13 (4.13)*
Satimage	**95.29 (0.53)**	95.22 (0.55)	95.22 (0.55)	95.23 (0.55)	*95.28 (0.53)*	95.22 (0.55)
Segmentation	91.17 (1.21)	91.66 (1.17)	91.71 (1.16)	**91.76 (1.17)**	91.17 (1.22)	91.67 (1.17)
Mammographic	*82.38 (3.37)*	*82.36 (3.67)*	*82.37 (3.69)*	*82.37 (3.69)*	**82.42 (3.39)**	*82.37 (3.66)*
Credit-approval	*86.29 (2.38)*	*86.52 (2.41)*	*86.49 (2.41)*	*86.5 (2.4)*	*86.28 (2.38)*	**86.52 (2.41)**
Ozone	*96.87 (0.79)*	*96.87 (0.79)*	*96.87 (0.79)*	*96.87 (0.79)*	*96.87 (0.79)*	**96.87 (0.79)**
Tic-tac-toe	86.23 (2.29)	84.97 (2.67)	85.09 (2.62)	85.28 (2.56)	**86.33 (2.34)**	85.01 (2.64)
Ilpd	*70.95 (3.31)*	*71.07 (3.44)*	*71.08 (3.44)*	*71.08 (3.44)*	*70.91 (3.28)*	**71.08 (3.44)**

Table 6.6 Summary performance statistics of the *A Priori* FM, the Uriz FM and the Sugeno λ-measure based ensemble classifiers with mixed classifiers. For clarity, the best classifiers are bold, the classifiers performances which are statistically significantly indifferent than the best are in *italics* and the ones which are statistically significantly worse than the best are underlined.

Datasets	Sugeno CFI	Sugeno RAV	Uriz CFI	Uriz RAV	*A Priori* CFI	*A Priori* RAV
Dermatology	95.59 (2.33)	**96.69 (1.87)**	*96.68 (1.88)*	*96.61 (1.87)*	95.22 (2.49)	*96.62 (1.9)*
Wine	*95.08 (5.87)*	*95.49 (4.05)*	94.59 (7.81)	94.38 (10.18)	*95.3 (5.81)*	**95.49 (4.05)**
Ecoli	95.22 (2.5)	*95.84 (2.63)*	**95.9 (2.62)**	*95.84 (2.64)*	95.37 (2.38)	*95.82 (2.62)*
Glass	91.52 (5.21)	**92.23 (5.21)**	*91.59 (5.98)*	89.45 (15.09)	91.45 (5.38)	*92.16 (5.03)*
Sonar	77.44 (6.96)	*81.28 (5.46)*	*81.42 (5.51)*	*80.56 (7.74)*	78.09 (6.78)	**81.56 (5.18)**
Ionosphere	**91.52 (4.24)**	90.61 (3.78)	90.83 (3.67)	*91.1 (3.84)*	*91.51 (4.07)*	90.7 (3.66)
SPECTF Heart	*79.7 (4.92)*	**80.09 (4.42)**	*79.78 (4.21)*	*79.54 (4.67)*	*79.57 (4.77)*	*79.83 (4.18)*
Bupa	66.97 (6.1)	*70.07 (5.4)*	**70.13 (5.74)**	*69.77 (5.9)*	66.74 (6.17)	*69.99 (5.55)*
WDBC	94.99 (2.58)	**96.41 (2.08)**	*96.37 (2.08)*	*96.38 (2.1)*	94.92 (2.68)	*96.37 (2.1)*
Haberman	67.95 (5.22)	*73.15 (4.56)*	*72.9 (4.55)*	*73.21 (4.54)*	68.29 (5.73)	**73.24 (4.56)**
Pima	75.58 (3.44)	**76.68 (3.18)**	*76.55 (3.14)*	*76.59 (3.12)*	75.48 (3.56)	*76.5 (3.15)*
Australian	84.35 (2.66)	*86.32 (2.62)*	*86.4 (2.47)*	*86.3 (2.52)*	84.28 (2.77)	**86.34 (2.56)**
SA Heart	68.17 (4.44)	*71.3 (4.15)*	70.7 (3.97)	*71.14 (4.38)*	68.26 (4.8)	**71.35 (4.12)**
Satimage	94.3 (0.81)	*95.25 (0.61)*	*95.22 (0.59)*	**95.25 (0.59)**	94.23 (0.81)	*95.24 (0.59)*
Segmentation	92.71 (1.87)	94.89 (1.35)	*95.21 (1.44)*	**95.42 (1.47)**	92.87 (1.99)	*95.05 (1.39)*
Mammographic	80.5 (3.45)	*82.14 (4.29)*	*81.99 (4.08)*	**82.62 (3.91)**	80.65 (3.41)	*82.35 (4.18)*
Credit-approval	84.67 (3.12)	*86.77 (2.85)*	*86.81 (2.91)*	**86.84 (2.83)**	84.52 (3.15)	*86.71 (2.9)*
Ozone	96.02 (0.89)	*96.8 (0.83)*	*96.78 (0.78)*	**96.81 (0.84)**	96.05 (0.82)	*96.8 (0.84)*
Tic-tac-toe	90.67 (2.77)	89.29 (2.34)	90.5 (2.32)	91.56 (2.4)	**92.13 (2.4)**	89.8 (2.4)
Ilpd	66.78 (4.45)	*70.01 (3.63)*	69.47 (3.61)	*70 (3.87)*	66.53 (4.46)	**70.14 (3.57)**

base classifier was also compared with DeFIMKL, MJSVM, Adaboost with trees, Bagging, and Random Forest for the same 20 UCI datasets.

The results show that the specific instance of an *A Priori* FM put forward in this case study is a robust way to construct FMs which perform well (with different aggregation operators) in the context of ensemble classification. Results highlight the strong value of leveraging available information, rather than relying on FM generation approaches which focus solely on the densities (such as the Sugeno λ-measure) and the potential of alternative FM generation approaches which do not rely on training the FM in respect to one specific aggregation operator. The results suggest that the proposed method to construct the *A Priori* FM leads to a more intuitive classification and performs well and efficiently.

To estimate how well the A Priori FM based ensemble classifiers perform, they also need to be compared with stat-of-the-art ensemble methods including DeFIMKL, MJSVM, Adaboost with trees, Bagging, and Random Forest. The next case study presents this comparison, and to be consistent the same 20 UCI datasets are chosen.

6.3 Case Study 3: Comparison of Aggregation Operator-Fuzzy Measure based Ensemble Algorithms with State-of-the-art Ensemble Algorithms

6.3.1 Problem Overview

The above two case studies concluded that *A Priori* FM based ensemble classifiers have performed well, however, their performances have only been compared to other aggregation operator-FM based ensemble classifiers. This leads to a research gap in literature i.e. there is no in-depth comparison of aggregation operator-FM based ensemble classifiers with traditional ensemble methods in the literature [23, 94, 91, 85, 86, 93]. In this case study we go a step further, and compare the proposed *A Priori* FM based ensemble classifier with traditional state-of-the-art ensemble classifiers such as Adaboost, Bagging, Random Forest and Majority Voting, using the same 20 UCI datasets from the previous case study. For a complete study, the DeFIMKL, a state-of-the-art aggregation operator based decision level fusion ensemble classifier, was also selected for comparison. This comparison is necessary to present a complete study on aggregation operator based ensemble classifiers resulting in better understanding. This comparison is necessary to present a complete study on aggregation operator based ensemble classifiers resulting in better understanding.

6.3.2 Background and Motivation

In the past decade, researchers have turned their attention towards aggregation operator-FM based ensemble classifiers and proposed a number of ensemble algorithms generating FMs from fuzzy densities, i.e. algorithmic FM [82, 81, 100]. For example, Wang et al. [82] proposed the use of posterior probabilities to obtain the fuzzy densities from the ensemble of heterogeneous classifiers. Subsequently, the Sugeno $\lambda-$measure was used to obtain the FM from the densities followed by aggregation using the CFI. The ensemble model was compared with five individual classifiers on the Satimage dataset. In another work, Fakhar et al. [81] proposed the use of the training accuracy and the fuzzy entropy (reliability of the information provided by each information source) to generate fuzzy densities followed by aggregation using the CFI. The proposed ensemble model was compared with seven fuzzy set theory based fusions, all of which are multi-biometric identification systems [163]. The accuracy of the proposed CFI-FM based ensemble outperformed the previously used classification models on the NIST database. Similarly, Wang et al. [100] proposed an aggregation operator based ensemble classification algorithm where the fuzzy densities are generated using the accuracy rate, error distance and the failure extent of the Neural Network models. The model was tested on the JAFFE facial expression database and compared with five Neural Network models.

In another set of studies, Anderson et al. [164] proposed the use of Genetic Algorithm (GA) (optimisation method) to learn FMs. This study indicated that aggregation operator based ensemble classifiers could learn the FMs from the training dataset, leading to efficient data-driven FMs. Hu et al. [94] extended the work and proposed a Fuzzy Integral-Genetic Algorithm (FIGA) ensemble classifier which aggregates results of SVM classifiers using the CFI w.r.t. FM learned through the hybrid of Sugeno λ-measure and GA. FIGA generated the initial measure through the use of Sugeno λ-measure followed by GA to search for an optimal FM through the error optimisation. The resultant ensemble was compared with the Multiple Kernel Learning Group Lasso (MKLGL) ensemble classifier on three datasets. FIGA performed better than MKLGL on all the three datasets.

Pinar and colleagues [23, 91, 85, 86, 93] built upon the previous works on data-driven FMs and proposed the Decision-level Fuzzy Integral Multiple Kernel Learning (DeFIMKL) algorithm as an alternative to algorithmic and algorithm-optimisation hybrid FMs, which aggregates the outputs of SVM classifiers through the use of CFI with respect to a FM learned through a regularised quadratic programming approach. Upon initial investigation in [23, 94] (where the DeFIMKL was compared to FIGA, MKGL and other FI-FM based

ensemble classifiers on six datasets), the accuracy of DeFIMKL based ensemble classification method was found to be better than classifiers based on multiple kernel learning. These works further concluded that the FIGA and the DeFIMKL were best among the feature level fusion classifiers and decision level fusion classifiers respectively.

The researchers in all the above works concluded that no comparison of the aggregation operator-FM based ensemble classifier was made with other state-of-the-art ensemble classification algorithms. Therefore, to contextualise the performance of the aggregation operator based ensemble classifiers, in the next section we present an in-depth study comparing the proposed *A Priori* FM based ensemble classifier and DeFIMKL with Adaboost, Bagging, Random Forest and MJSVM (Majority Voting with SVM) for 20 datasets from UCI datasets.

6.3.3 Experiments

The Adaboost ensemble was run with 200 decision trees, Bagging also with 200 but Random Forest ensemble method with 100 decision trees. The ensemble size is chosen based on experiments. It is important to note that not much discussion is given to determine the size of these ensembles, which is one of the major underlying issue within ensemble classification algorithms, as they are not in the scope of the case study. The DeFIMKL and the MJSVM ensemble methods used five Support Vector Machine (SVM) algorithms with Radial Basis Function (RBF) kernels as their base classifiers. The five RBF kernels with their width (σ) evenly spaced between $0.5 - 1.5/(number\ of\ features)$ were used for both the ensemble methods. Additionally, $L1$ regularisation with $\lambda = 0.5$ was used for all the datasets. The focus of this work is to show the effect on the final output when different types of aggregation models are employed (e.g. DeFIMKL and MJSVM), and thus the settings for the RBF kernels and other ensemble methods were kept the same for all the datasets.

6.3.4 Results and Discussion

Table 6.7 reports the average accuracies of the A Priori FM-CFI based ensemble classifier, the A Priori FM-RAV based ensemble classifier, DeFIMKL, MJSVM, Adaboost, Bagging and Random Forest ensemble classification algorithms with standard deviations over 100 runs. In each trial, 80% of the data were randomly used for training the classifiers and the remaining 20% for testing. A series of two-sample t-tests were conducted, which compared the accuracy of each algorithm against that of the highest performing algorithm for each dataset. To illustrate the results of these tests, both the absolute highest performing algorithm, along with any further algorithms that were found not to have a significantly lower (at

Table 6.7 Summary performance statistics of the *A Priori* FM based ensemble classifier comparison with state-of-the-art ensemble classifiers. For clarity, the best classifiers are bold, the classifiers performances which are statistically significantly indifferent than the best are in *italics* and the ones which are statistically significantly worse than the best are underlined.

Datasets	*A Priori* CFI	*A Priori* RAV	DeFIMKL	MJSVM	Adaboost with trees	Bagging	Random Forest
Dermatology	96.81 (2.16)	97.15 (1.97)	*97.35 (1.86)*	**97.47 (1.75)**	96.69 (2.05)	*97.31 (1.77)*	95.47 (2.65)
Wine	98.86 (2)	*99.16 (1.66)*	**99.44 (1.18)**	*99.42 (1.44)*	96.81 (2.91)	97.78 (2.68)	96.44 (3.14)
Ecoli	**96.88 (2.1)**	*96.71 (2.13)*	*96.77 (1.84)*	*96.84 (1.85)*	95.57 (2.01)	96.34 (2.36)	95.82 (2.29)
Glass	92.93 (3.95)	92.43 (4.32)	*94 (3.8)*	*94.16 (3.73)*	*93.91 (3.92)*	**94.28 (3.13)**	92.72 (3.78)
Sonar	80.51 (5.99)	80.44 (6.34)	**84.57 (4.7)**	*83.76 (5.4)*	*83.17 (5.89)*	*83.21 (6.2)*	79.31 (6.2)
Ionosphere	*94.69 (2.66)*	94.13 (2.61)	**94.61 (2.71)**	*94.34 (2.59)*	90.41 (3.72)	92.76 (2.74)	91.15 (3.57)
SPECTF Heart	*80.11 (5.19)*	*80.06 (5.19)*	79.19 (4.4)	79.48 (4.18)	79.78 (4.98)	**81.85 (5.01)**	*80.81 (4.23)*
Bupa	67.36 (5.23)	67.75 (5.34)	*69.88 (4.95)*	*69.77 (4.71)*	**70.93 (5.32)**	*70.43 (5.02)*	*69.77 (5.65)*
WDBC	96.7 (1.79)	96.83 (1.78)	*97.22 (1.64)*	**97.28 (1.62)**	96.59 (2.03)	95.51 (1.94)	95.11 (2.15)
Haberman	*73.05 (4.68)*	*73.21 (4.81)*	*73.77 (4.84)*	**73.92 (4.88)**	*73.4 (4.25)*	69.34 (4.99)	69.24 (4.52)
Pima	*75.95 (2.99)*	*75.95 (2.83)*	*76.12 (3.17)*	*76.49 (3.14)*	*75.97 (3.22)*	**76.61 (3.16)**	*75.97 (3.21)*
Australian	84.75 (3.01)	85.14 (2.94)	85.75 (2.42)	85.85 (2.35)	85.94 (2.79)	**86.96 (2.68)**	*86.37 (2.81)*
SA Heart	*70.95 (4.08)*	*71.13 (4.13)*	*71.16 (4.04)*	**71.66 (3.89)**	69.31 (4.23)	69.39 (4.23)	68.63 (4.07)
Satimage	95.28 (0.53)	95.22 (0.55)	**95.69 (0.58)**	95.53 (0.56)	94.09 (0.6)	95.6 (0.47)	95.17 (0.69)
Segmentation	91.17 (1.22)	91.67 (1.17)	91.81 (1.15)	91.46 (1.08)	92.86 (1.53)	*96.03 (1.1)*	**96.29 (1.26)**
Mammographic	*82.42 (3.39)*	*82.37 (3.66)*	*82.16 (3.65)*	**82.88 (4.02)**	*81.8 (5.71)*	78.6 (5.39)	77.21 (5.02)
Credit-approval	86.28 (2.38)	86.52 (2.41)	**86.89 (2.38)**	*86.66 (2.73)*	86.47 (2.84)	*87.4 (3.15)*	*87.02 (2.83)*
Ozone	96.87 (0.79)	96.87 (0.79)	**97.12 (0.79)**	*97.1 (0.8)*	96.76 (0.77)	96.85 (0.85)	96.64 (0.83)
Tic-tac-toe	86.33 (2.34)	85.01 (2.64)	89.64 (2.25)	88.54 (2.39)	84.34 (2.5)	94.38 (1.85)	**95.35 (1.95)**
Ilpd	70.91 (3.28)	71.08 (3.44)	**72.18 (3.58)**	*71.79 (3.58)*	68.18 (4.49)	68.01 (4.38)	67.77 (4.15)

$p<.05$) accuracy, are highlighted in bold. Thus, the values highlighted in bold represent the algorithm(s) with the highest accuracy for each dataset and will include all values that were not found to significantly differ from the best-performing model.

From Table 6.7 it can be observed that the DeFIMKL and MJSVM were the overall *best* classifiers, showing good performance in 16 and 15 datasets respectively. It can be observed that DeFIMKL is either equivalent or better in performance than MJSVM. However, MJSVM has an advantage as it reduces the memory complexity without having to learn the huge 2^n FM vector i.e. taking the majority voting of the SVMs results in similar performance to the non-linear CFI aggregation of the SVMs. The other classifiers: CFI with *A Priori* FM, RAV with *A Priori* FM, Adaboost with trees, Bagging and the Random Forest had the *best* accuracies in 7, 7, 6, 9 and 7 datasets respectively. DeFIMKL is the overall *best* aggregation operator based ensemble classifier, showing that superior performance can be obtained by optimising the FM in respect to a specific operator (the CFI in this case). However, it is important to understand that the *A Priori* measures are designed for data fusion problems. Moreover, the *A Priori* FM based classifiers do not employ optimisation towards a specific aggregation operator, and still able to achieve good performance across a number of datasets. Overall, the results from the Table 6.7 indicate that aggregation operator-FM based ensemble classification algorithm (especially, DeFIMKL) can be used in applicaitons using traditional ensemble algorithms.

6.3.5 Summary

The aggregation operator-FM based ensemble classifiers have been used in various applications over the past decade, yet in most papers, a dataset oriented comparison with state-of-the-art ensemble classification has been made. In this case study we address this gap and compare the proposed *A Priori* FM driven ensemble classifier (for both the CFI and RAV aggregation operators) and DeFIMKL (the state-of-the-art aggregation operator based ensemble classifier) algorithms with tradition ensemble classifiers including Adaboost, Bagging, MJSVM and Random Forest. The experiments are conducted on 20 datasets from the UCI machine learning repository.

On comparing the accuracies, it can be observed that DeFIMKL and MJSVM have the best performances. The CFI with *A Priori* FM, RAV with *A Priori* FM, Adaboost with trees, Bagging and Random Forest were best in 7, 7, 6, 9 and 7 datasets respectively. Although, the *A Priori* FM driven ensemble algorithms did not perform as well as DeFIMKL, it is important to note an important difference between the two i.e. the *A Priori* FM based classifiers do

not employ optimisation w.r.t. a specific aggregation operator. Note: Good results for the DeFIMKL approach shows that a superior performance can be obtained by optimising the FM in respect to a specific operator (the CFI in this case).

It is important to understand that the *A Priori* FMs are designed for data fusion problems, such as the one presented in case study 1. Moreover, the *A Priori* FM are useful in understanding the fusion step in the ensemble (case study 1), the reason for which they were designed. Lastly, it can be inferred that the non-linear aggregation can squeeze out a better performance at the cost of additional memory complexity, although a detailed study is needed to fully answer this question, which is out of scope for this thesis.

6.4 Overall Summary

The current approaches to capture the FMs can be classified into three broad categories: Expert-driven, Optimisation and Algorithm-driven. Each category suffers from some drawback. With an increase in the number of information sources, expert-driven FMs become virtually impossible to define as the factors to be learned increase exponentially. Although optimisation algorithms work well in a number of applications (specifically the ones with training data and labels), they need to learn the FM with respect to the selected aggregation operator. Lastly, the Algorithm based FMs do not capture the interaction among the information sources.

A new approach employing external knowledge to construct FM is proposed (in Chapter 5), called the *A Priori* FM. An Algorithm 2 was also proposed to construct these external knowledge to the *A Priori* FM. This new FM can be defined for a large number of information sources, can capture the interaction among these sources and be used by all the aggregation operators without needing to be defined individually for each operator. The external information to construct the FM can come from various sources, and in this thesis we put forward two approaches to capture them: 1) Classifier-performance driven, and 2) Literature knowledge driven. We also developed one specific instance of the A Priori FM: one which captures the quality of individual classifiers (and their combinations) in order to then enable a fusion-based ensemble classifier.

To evaluate this new *A Priori* FM, we present three detailed case studies employing the *A Priori* FM driven ensemble classifier across multiple datasets. In the first case study, the *A Priori* FM driven ensemble classification was employed on a multi-region face recognition application. The second case study builds upon the previous one to present a more detailed analysis of the *A Priori* FM. Across both the studies, we compare the performance of *A Priori*

FM with state-of-the-art methods Sugeno λ-measure and the Uriz FM; for both the CFI and the RAV aggregation operators. To complete the analysis, the *A Priori* FM driven ensemble classifier was also compared with a state-of-the-art aggregation operator-FM based ensemble classifier *DeFIMKL* and four traditional ensemble classifiers *Adaboost*, *Bagging*, *Random Forest* and *Majority Voting*. Overall, the three empirical case studies present a complete study on *A Priori* FM.

It was observed that the proposed *A Priori* FM performs well with both the CFI and the RAV operator. This study was also useful in understanding the configurations of the fusion step in the ensemble i.e. interpretable ensemble of classifiers. From these case studies it was also inferred that the use of external *a priori* knowledge can improve classification, leading to a better, interpretable and robust (can be used across multiple datasets) ensemble model.

Chapter 7

Conclusions

7.1 Contributions

New Classification Framework for difficult to classify data

One challenge within the applications using unlabelled datasets was to generate a robust ensemble classification method capable of classifying data points which are difficult to label. Chapter 3 in this thesis introduced a robust classification framework by adding an '*ensemble classification*' step after the ensemble clustering step. In doing so, this thesis improved the classification performances within the unlabelled datasets. Subsequently, Chapter 4 presented its application on multiple datasets, concluding that the classification was able to achieve a real-world impact. The classification framework has contributed to publications P1 and P2 listed in dissemination section, achieving objectives 1, 2 and 4 listed in Section 1.2 of Chapter 1.

HGV Driving Stereotype Identification

To determine the driving stereotypes within the U.K. HGV community, the two-stage classification framework is employed on a telematics dataset containing the driving incidents of the drivers. Eight novel driving stereotypes were uncovered which are now the state-of-the-art in this domain (more details in Chapter 4 and publication P3 in the dissemination section), achieving objectives 4 from Section 1.2 of Chapter 1. Moreover, these findings assist policy makers and stakeholders in the HGV industry to better understand their drivers and to define actions to improve road economy and safety.

External Information driven *A Priori* FM

One objective of the thesis was to develop new approaches for classifier fusion across the cases where additional data about the information sources is available. Thus, in Chapter 5 we introduced the novel *A Priori* FM which leverages external information in order to drive ensemble classification, achieving objective 3 listed in Section 1.2 of Chapter 1. A comprehensive study introducing a decision level fusion ensemble classifier based on this *A Priori* FM was also presented. Its application on multiple classification applications was then presented in Chapter 6. This contribution has resulted in publication P4 (and a journal paper currently under preparation).

In-depth Comparison of Aggregation Operator-Fuzzy Measure based Ensemble Algorithms with the State-of-the-art Ensemble Algorithms

While the aggregation operator based ensemble classifiers have been compared to other similar classifiers and a number of individual classifiers, it has thus far not been extensively compared to the state-of-the-art ensemble classification methods. Thus, a detailed comparison of the aggregation operator-FM based ensemble methods, specifically the *A Priori* FM based classifier and the DeFIMKL algorithm, is made with state-of-the art ensemble methods including Adaboost, Bagging, Random Forest and Majority Voting. Chapter 6 presents this comparison, achieving objective 5 listed in Section 1.2 of Chapter 1 and resulting in publication P5.

7.2 Thesis Summary

In pattern recognition, classification of data is of prime importance. Ensemble classifiers, which are models that combine results of multiple individual classifiers, have become a state-of-the-art in this domain. Although the ensemble classification models have been employed across a wide range of applications, some key challenges still remain in order to generate robust models. In particular, two challenges are of interest in this thesis; the first is classification of data points which are difficult to classify and the second is to construct an improved decision level fusion ensemble classifier by leveraging additional external information. This thesis explores ways of addressing the two aforementioned challenges.

The core aim of this thesis is to improve existing mechanisms underpinning ensemble classification algorithms and to develop new robust (i.e. a general framework that can adapt

to different dataset) methods for efficient ensemble classification. In order to achieve this aim, five objectives have been identified, detailed in Section 1.2 of Chapter 1.

Chapter 2 outlines an overview of background material and existing literature that are used and referred to in this thesis. A review of the current state-of-the-art approaches for ensemble classification is presented. The review includes traditional methods like Adaboost, Bagging and recently developed methods like aggregation operator and Fuzzy Measure (based ensemble classifiers). The chapter also highlights the current (aforementioned) research gaps in constructing a robust ensemble classification approach. Lastly, the applicaitons employed to evaluate the new methods proposed in this thesis are also reviewed.

One key issue within the unlabelled datasets is classification of data points which are difficult to classify. Recently, a series of works have leveraged an ensemble of multiple clustering algorithms to elucidate the most characteristic groups across these unlabelled datasets [8, 115]. Due to limitations of these ensemble clustering methods, a number of data remained unclassified, leading to incomplete representation of the dataset. In Chapter 3 the problem of 'difficult to classify data' is demonstrated using the Nottingham Breast Cancer (unlabelled dataset) as a case study. To tackle this challenge we introduce a novel two stage robust classification framework, which couples an ensemble classification stage after an ensemble clustering stage (leading to Contribution 1 discussed in Section 7.1). In doing so, we fulfil the first two objectives of the thesis: 1) investigating *difficult to classify data points* within the classification applications employing unlabelled datasets for a robust classification approach, and 2) establishing an enhanced and robust classification framework to tackle 'difficult to classify points' within the unlabelled datasets. The classification framework can identify core groups in an unlabelled dataset, distribute data in the identified groups and improve the final classification by assigning unclassified data to one of the groups.

Chapter 4 empirically evaluates the two stage classification framework introduced in the previous chapter. Three comprehensive empirical studies are presented in this chapter, which evaluates the robustness and real-world impact of the proposed classification framework, addressing 'Objective 4' of the thesis. The three case studies include biological group stratification in the Edinburgh Breast Cancer Dataset, driving stereotypes identification in the Microlise's HGV dataset[1] and characteristic groups stratification within multiple standard datasets from UCI repository. The results on these datasets illustrate that by using the novel classification framework, an improved, reliable and robust (a method that can be employed and adapt across different datasets) classification is obtained; which is verified by statistical

[1]Note: The application of classification framework on the HGV dataset leads to eight state-of-the-art driving stereotypes i.e. Contribution 2 from Section 7.1

tests, visualisation techniques, cluster quality assessment and interpretation from experts (ground truth in case of UCI repository).

Chapter 5 moves beyond Chapters 3 and 4 and looks at a different aspect of ensemble classification: combining classifier outcomes for the cases where additional data about the information sources is available. In doing so it fulfils 'Objective 3' of the thesis. One approach to capture this insight from data is aggregation operators which can capture different levels of knowledge and different knowledge content by evaluating interactions among different information sources. In this thesis we use two aggregation operators, the state-of-the-art CFI and the recently introduced RAV. Aggregation operators can combine informations from different sources with respect to a FM. FMs are bounded monotonic capacities which can capture the worth of both the individual sources and all their possible combinations. Current FMs suffer from some drawbacks or cannot leverage the external information, therefore this chapter introduced a fourth category to construct the FMs, the *A Priori* FM. This method converts the external information (*a priori* knowledge) on the relevance of different combinations of sub-sources to construct the A Priori FM. We examine two different pathways to capture the *a priori* knowledge: 1) Classifier (Machine Learning) driven, and 2) Literature knowledge driven. To support the decision level fusion step, the ensemble classification algorithm is developed by employing the *A Priori* FMs generated from the two aforementioned knowledge sources, with both the CFI and the RAV aggregation operators. Lastly, in this chapter we compare the two aggregation operators in order to determine the optimal aggregation operator for *A Priori* FM. The comparison results suggested that both the operators could be employed with this new FM. In summary, the chapter presents a new approach to construct the FM (i.e. the *A Priori* FM) which leverages external information which could be employed with both the CFI and the RAV aggregation operators, leading to Contribution 3 from Section 7.1.

Chapter 6 presents three case studies to empirically evaluate the *A Priori* FM introduced in Chapter 5. In the first case study we use multi-region facial recognition application to demonstrate the performance of *A Priori* FM, employing both the CFI and the RAV as aggregation operators. We also evaluate its performance by contrasting it with the Uriz FM and the Sugeno λ-measure; again for both the CFI and the RAV as aggregation operators. A more comprehensive study was essential to comprehend the performance of this new *A Priori* FM and subsequently, in the second case study we compare its performance with Uriz FM and Sugeno λ-measure on 20 datasets from the UCI machine learning repository. Lastly, in the third study the *A Priori* FM based ensemble classifier is contextualised to the wider literature and compared with state-of-the-art aggregation operator-FM based ensemble

classifier DeFIMKL and traditional ensemble classifiers Adaboost, Bagging, Random Forest and Majority Voting. The results on the selected datasets suggested that the aggregation operator based ensemble classifier performs both well and efficiently, indicating that it is a viable alternative when selecting ensemble classifiers. The results from the three case studies suggested that the use of external information based FM (*A Priori* FM) and the aggregation operator based ensemble classifier performs both well and efficiently, indicating that it is a viable alternative when selecting ensemble classifiers. The results also indicated that the non-linear fusion of decision level outputs offered by the aggregation operators provided details of the fusion step in the ensemble i.e. interpretability, which was useful in improving and understanding the overall classification outcomes. In summary, the chapter fulfils 'Objective 4 and 5' of the thesis i.e. real-world impact of the new *A Priori* FM based classification approach and resulting in Contribution number 4 presented in Section 7.1 i.e. a comprehensive study contextualising FM driven ensemble classifiers against a wider literature.

In can be concluded from the chapters that this thesis achieves the aim of improving existing mechanisms underpinning ensemble classification algorithms (by introducing a novel two stage classification framework) and also develops a new robust method for efficient ensemble classification (A Priori FM based ensemble classification). As part of the aim the thesis is also able to fulfil the five objectives. In the following sections, we discuss the limitations of the research in this thesis, possible directions for future works, contributions of this thesis and the list of publications.

7.3 Limitations

Overfitting the results of the two stage classification framework

Chapter 3 presents a two stage ensemble clustering-ensemble classification framework to improve the classification (of difficult to classify datasets) across unlabelled datasets. Chapter 4 empirically evaluates the performance of this classification framework by presenting is application on three case studies. Although, superior classification performance is observed across these studies, the algorithms in the framework might suffer from the problem of overfitting (however, the performance results on UCI datasets might suggest that overfitting is not a limitation). The framework is also not able to achieve the highest accuracy in comparison to other supervised classification algorithm accuracies in literature, however, these differences are small and acceptable.

Reliance on experts due to lack of ground truth

The two case studies of the classification framework (in Chapter 4) on the Edinburgh Breast Cancer dataset and the HGV dataset are limited by the fact that there is no ground truth to validate the findings observed in this thesis. Therefore, there is reliance on experts to verify the results.

Reliance on external factors in HGV case study

In case of HGV dataset, the inputs related to road characteristics, weather conditions and traffic conditions, are not included explicitly in the analysis, which themselves might be causes of the driving incidents. With our current data, it is not possible to distinguish the driving incidents caused by these external factors. However, we do believe that by considering driving data across the whole year under different conditions, and considering several roads in different locations across the UK, the impact of those external factors is overall reduced.

Multi-class A Priori FM based ensemble classifiers

Chapter 5 presents an *A Priori* FM based ensemble classifier which is capable of leveraging external information for decision level fusion ensemble classification. Chapter 6 presents a comprehensive study of its application across multiple datasets and its detailed comparison with other state-of-the-art ensemble methods. Although the *A Priori* FM based ensemble classifier performed well and is able to its objective, a binary classifier is constructed using it. A multi-class ensemble classifier still needs to be developed.

7.4 Future Works

Core driving profiles for real-time classification

As future directions, we intend to merge the dataset from the year of 2015 with data from 2014, 2016 and 2017 to investigate whether stereotypes have changed over the years and whether their proportions change in different seasons. We then aim to generate a real-time classification system using these core driving profiles. We also intend to include other types of incidents in the analysis, such as harsh cornering, which seem to be relevant for HGV and cargo safety.

Application of *A Priori* FM in Computer Vision

As part of an expanded work on *a priori* FMs, we have developed *a priori* FMs which leverage external information from the images (foot, ear and face) of the babies to support improved fusion for gestational age prediction, showing a pathway for delivering strong information aggregation using interpretable and validatable FMs, with rich scope for further research.

Additional analysis of the *A Priori* FM

In the future, we will focus on exploring the robustness of *a priori* FMs when employed with a wider set of aggregation operators, targeting the Sugeno FI and the RAV operator for different values of p. Further, we aim to develop *a priori* FMs which can leverage information from other potential external sources to support improved fusion, showing a pathway for delivering strong information aggregation using interpretable and validatable FMs. As previously mentioned in the limitation section, the *A Priori* FM based ensemble classifier is a binary classifier, and thus, in future we aim to present a detailed study for multi-class classification.

Employ *A Priori* FM based ensemble classifier in two stage classification framework

In Chapters 3 a two-stage classification framework, employing ensemble classifiers stage followed by ensemble clustering stage, is introduced. Chapter 5 presents a novel *A Priori* FM based ensemble classifier. Intuitively, one future work would be to employ the presented *A Priori* FM based ensemble classifier as one of the ensemble models in the classification framework.

References

[1] Complete Guide to Deep Neural Networks – Part 1. https://pythonmachinelearning.pro/complete-guide-to-deep-neural-networks-part-1. last visited: 16/09/2019.

[2] What is an artificial neural network? https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network. last visited: 16/09/2019.

[3] L.L. Hong, W.C. Heng, Y.T. Fui, and K.H. Meian. A review of nearest neighbor-support vector machines hybrid classification models. *Journal of Applied Sciences*, 10:1841–1858, 2010.

[4] Introduction to Support Vector Machines. https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html. last visited: 16/09/2019.

[5] Thomas G Dietterich. Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1857:1–15, 2000.

[6] Ludmila I Kuncheva. A bound on kappa-error diagrams for analysis of classifier ensembles. *IEEE Transactions on knowledge and data engineering*, 25(3):494–501, 2013.

[7] Christian Wagner, Timothy C Havens, and Derek T Anderson. The Arithmetic Recursive Average as an Instance of the Recursive Weighted Power Mean. In *IEEE Int. Conf. Fuzzy Systems. 2017*, 2017.

[8] Daniele Soria, Jonathan M. Garibaldi, Federico Ambrogi, Andrew R. Green, Des Powe, Emad Rakha, R. Douglas Macmillan, Roger W. Blamey, Graham Ball, Paulo J G Lisboa, Terence A. Etchells, Patrizia Boracchi, Elia Biganzoli, and Ian O. Ellis. A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. *Computers in Biology and Medicine*, 40(3):318–330, 2010.

[9] A R Green, D G Powe, E A Rakha, D Soria, C Lemetre, C C Nolan, F F T Barros, R D Macmillan, J M Garibaldi, G R Ball, and I O Ellis. Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers. *British journal of cancer*, 109(7):1886–94, 2013.

[10] 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. In *Artificial Intelligence in Medicine*, volume 58, pages 175–184, 2013.

[11] AT&T Dataset. https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. last visited: 18-05-2019.

[12] Zoran Constantinescu, Cristian Marinoiu, and Monica Vladoiu. Driving style analysis using data mining techniques. *Int J of Comp Communications & Control*, 5(5):654–663, 2010.

[13] Hai Long Nguyen, Yew Kwong Woon, and Wee Keong Ng. A survey on data stream clustering and classification. *Knowledge and Information Systems*, 45(3):535–569, dec 2015.

[14] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

[15] Mariana Belgiu and Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.

[16] Ning Chen, Bernardete Ribeiro, and An Chen. Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 45(1):1–23, 2016.

[17] Scott Krig. *Computer vision metrics*. Springer, 2016.

[18] Sandhya Samarasinghe. *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. Auerbach publications, 2016.

[19] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[20] Saba Bashir, Usman Qamar, and Farhan Hassan Khan. Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. *Quality and Quantity*, 49(5):2061–2076, 2014.

[21] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

[22] Grazziela P. Figueredo, Philip R. Quinlan, Mohammad Mesgarpour, Jonathan M. Garibaldi, and Robert I. John. A Data Analysis Framework to Rank HGV Drivers. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2015-Octob:2001–2006, sep 2015.

[23] Anthony J. Pinar, Joseph Rice, Lequn Hu, Derek T. Anderson, and Timothy C. Havens. Efficient Multiple Kernel Classification using Feature and Decision Level Fusion. volume 6706, 2016.

[24] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25, 2017.

[25] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2):47–58, 2013.

[26] Abraham Pouliakis, Efrossyni Karakitsou, Niki Margari, Panagiotis Bountris, Maria Haritou, John Panayiotides, Dimitrios Koutsouris, and Petros Karakitsos. Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future. *Biomedical engineering and computational biology*, 7:1–18, 2016.

[27] Robi Polikar. Pattern recognition. *Wiley Encyclopedia of Biomedical Engineering*, 4174(5):1–22, 2006.

[28] J Khan, J S Wei, M Ringnér, L H Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, C R Antonescu, C Peterson, and P S Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–9, 2001.

[29] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[30] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[31] P. J. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, 1974.

[32] Ahmet Mert, Niyazi Kılıç, Erdem Bilgili, and Aydin Akan. Breast Cancer Detection with Reduced Feature Set. *Computational and Mathematical Methods in Medicine*, 2015:1–11, 2015.

[33] DA Adeniyi, Z Wei, and Y Yongquan. Automated web usage data mining and recommendation system using k-nearest neighbor (knn) classification method. *Applied Computing and Informatics*, 12(1):90–108, 2016.

[34] M Sarkar and T Y Leong. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. *Proceedings of the AMIA Annual Symposium. American Medical Informatics Association*, pages 759–763, 2000.

[35] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, and Abdelkader Benyettou. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications*, 62(1):975–8887, 2013.

[36] Bernhard Schlkopf, Alexander J Smola, and Francis Bach. Learning with kernels: support vector machines, regularization, optimization, and beyond. 2018.

[37] Bernhard Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

[38] Min-Wei Huang, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. Svm and svm ensembles in breast cancer prediction. *PloS one*, 12(1):e0161501, 2017.

[39] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(4):463–484, jul 2012.

[40] Sangyoon Oh, Min Su Lee, and Byoung Tak Zhang. Ensemble learning with active example selection for imbalanced biomedical data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):316–325, mar 2011.

[41] A. Acharya, E. R. Hruschka, J. Ghosh, and S. Acharyya. C3E: A framework for combining ensembles of classifiers and clusterers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6713 LNCS:269–278, 2011.

[42] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78, 2012.

[43] Dewan Md Farid, Li Zhang, Alamgir Hossain, Chowdhury Mofizur Rahman, Rebecca Strachan, Graham Sexton, and Keshav Dahal. An adaptive ensemble classifier for mining concept drifting data streams. *Expert Systems with Applications*, 40(15):5895–5906, 2013.

[44] Rafal Lysiak, Marek Kurzynski, and Tomasz Woloszynski. Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers. *Neurocomputing*, 126:29–35, 2014.

[45] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, Dinani Amorim, and Dinani Amorim Fernández-Delgado. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

[46] David H. Wolpert. The Supervised Learning No-Free-Lunch Theorems. In *Soft Computing and Industry*, pages 25–42. Springer, London, 2002.

[47] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems*, pages 231–238, 1995.

[48] Ludmila I Kuncheva. That elusive diversity in classifier ensembles. In *Iberian conference on pattern recognition and image analysis*, pages 1126–1138. Springer, 2003.

[49] Amanda JC Sharkey and Noel E Sharkey. Combining diverse neural nets. *The Knowledge Engineering Review*, 12(3):231–247, 1997.

[50] Dragos D Margineantu and Thomas G Dietterich. Pruning adaptive boosting. In *ICML*, volume 97, pages 211–218, 1997.

[51] Ron Kohavi, David H Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. In *ICML*, volume 96, pages 275–83, 1996.

[52] Rafal Lysiak, Marek Kurzynski, and Tomasz Woloszynski. Probabilistic approach to the dynamic ensemble selection using measures of competence and diversity of base classifiers. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 229–236. Springer, 2011.

[53] Hamid Parvin, Miresmaeil MirnabiBaboli, and Hamid Alinejad-Rokny. Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence*, 37:34–42, 2015.

[54] Michał Woźniak and Manuel Graña. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014.

[55] Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. Ensemble learning for data stream analysis: a survey. *Information Fusion*, 37:132–156, 2017.

[56] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *journal of computer and system sciences*, 55:119–139, 1997.

[57] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[58] Quan Sun and Bernhard Pfahringer. Bagging Ensemble Selection. *Proceedings of the 24th international conference on Advances in Artificial Intelligence*, pages 251–260, 2011.

[59] V F Rodriguez-Galiano, B Ghimire, J Rogan, M Chica-Olmo, and J P Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67(1):93–104, 2012.

[60] Grazziela P Figueredo, Philip R Quinlan, Mohammad Mesgarpour, Jonathan M Garibaldi, and Robert I John. A data analysis framework to rank HGV drivers. In *2015 IEEE 18th Int Conf on Intelligent Transp Sys*, pages 2001–2006, 2015.

[61] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, 42(5):2785–2797, 2015.

[62] Sukhdev Singh Ghuman. Clustering techniques-a review. 2016.

[63] Arpita Nagpal, Arnan Jatain, and Deepti Gaur. Review based on data clustering algorithms. In *Information & Communication Technologies (ICT), 2013 IEEE Conference on*, pages 298–303. IEEE, 2013.

[64] Xiuge Wu, Tinghuai Ma, Jie Cao, Yuan Tian, and Alia Alabdulkarim. A comparative study of clustering ensemble algorithms. *Computers & Electrical Engineering*, 68:603–615, 2018.

[65] Tossapon Boongoen and Natthakan Iam-On. Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28:1–25, 2018.

[66] D Soria, J M Garibaldi, E Biganzoli, and I O Ellis. A Comparison of Three Different Methods for Classification of Breast Cancer Data. *2008 Seventh International Conference on Machine Learning and Applications*, pages 619–624, 2008.

[67] Grigorios Tzortzis and Aristidis Likas. The MinMax k-Means clustering algorithm. *Pattern Recognition*, 47(7):2505–2516, jul 2014.

[68] Utkarsh Agrawal, US Tiwary, Soumava Kumar Roy, and DS Prashanth. K-means clustering for adaptive wavelet based image denoising. In *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, pages 134–137. IEEE, 2015.

[69] M. Emre Celebi, Hassan A. Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, jan 2013.

[70] H S Park and C H Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2 PART 2):3336–3341, 2009.

[71] Alexander Strehl and Joydeep Ghosh. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[72] Joydeep Ghosh and Ayan Acharya. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315, 2011.

[73] Lior Rokach. Ensemble-based classifiers. *Artif Intell Rev*, 33:1–39, 2010.

[74] SANDRO VEGA-PONS and JOSÉ RUIZ-SHULCLOPER. A Survey of Clustering Ensemble Algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, may 2011.

[75] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesus M. Perez, and Inigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[76] F. H. C. Marriott. Practical problems in a method of cluster analysis. *Biometrics*, 27(3):pp. 501–514, 1971.

[77] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Com in Statistics-Simulation and Comp*, 3(1):1–27, 1974.

[78] Jean-Charles Lamirel, Nicolas Dugué, and Pascal Cuxac. New efficient clustering quality indexes. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3649–3657. IEEE, 2016.

[79] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):pp. 387–397, 1971.

[80] Anthony WF Edwards and L Luka Cavalli-Sforza. A method for cluster analysis. *Biometrics*, pages 362–375, 1965.

[81] Khalid Fakhar, Mohamed El Aroussi, Mohamed Nabil Saidi, and Driss Aboutajdine. Applying the upper integral to the biometric score fusion problem in the identification model. *2015 International Conference on Electrical and Information Technologies (ICEIT)*, 6(3):494–504, mar 2015.

[82] Qinghua Wang, Canliang Zheng, Hongtao Yu, and Donghua Deng. Integration of heterogeneous classifiers based on choquet fuzzy integral. *Proceedings - 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2015*, 1:543–547, aug 2015.

[83] Christian Wagner and Derek T Anderson. Extracting meta-measures from data for fuzzy aggregation of crowd sourced information. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2012.

[84] Luis Garmendia. The evolution of the concept of fuzzy measure. In *Intelligent Data Mining*, pages 185–200. Springer, 2005.

[85] Anthony J Pinar, Timothy C Havens, Muhammad Aminul Islam, and Derek T Anderson. Visualization and learning of the Choquet integral with limited training data. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, jul 2017.

[86] Anthony J. Pinar, Derek T. Anderson, Timothy C. Havens, Alina Zare, and Titilope Adeyeba. Measures of the Shapley index for learning lower complexity fuzzy integrals. *Granular Computing*, 2(4):303–319, dec 2017.

[87] Michio Sugeno. Theory of fuzzy integrals and its applications. *Doct. Thesis, Tokyo Institute of technology*, 1974.

[88] M. Uriz, D. Paternain, H. Bustince, and M. Galar. A first approach towards the usage of classifiers' performance to create fuzzy measures for ensembles of classifiers: a case study on highly imbalanced datasets. In *International Conference on Fuzzy Systems (FUZZ)*. IEEE, 2018.

[89] Michel Grabisch. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3):445–456, 1996.

[90] Toshiaki Murofushi and Michio Sugeno. A learning model using fuzzy measures and the Choquet integral. *Proc. of the 5th Fuzzy System Symposium*, 29:213–218, 1989.

[91] Derek T. Anderson, Stanton R. Price, and Timothy C. Havens. Regularization-based learning of the Choquet integral. In *IEEE International Conference on Fuzzy Systems*, pages 2519–2526. IEEE, jul 2014.

[92] Muhammad Aminul Islam, Derek T. Anderson, Anthony J. Pinar, and Timothy C. Havens. Data-Driven Compression and Efficient Learning of the Choquet Integral. *IEEE Transactions on Fuzzy Systems*, PP(99), 2017.

[93] Anthony Pinar, Timothy C. Havens, Derek T. Anderson, and Lequn Hu. Feature and decision level fusion using multiple kernel learning and fuzzy integrals. *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, aug 2015.

[94] Lequn Hu, Derek T. Anderson, and Timothy C. Havens. Multiple kernel aggregation using fuzzy integrals. *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, jul 2013.

[95] Ling Zhang, De Qun Zhou, Peng Zhou, and Qi Ting Chen. Modelling policy decision of sustainable energy strategies for Nanjing city: A fuzzy integral approach. *Renewable Energy*, 62:197–203, 2014.

[96] Derek T. Anderson, Timothy C. Havens, Christian Wagner, James M. Keller, Melissa F. Anderson, and Daniel J. Wescott. Extension of the fuzzy integral for general fuzzy set-valued information. *IEEE Transactions on Fuzzy Systems*, 22(6):1625–1639, dec 2014.

[97] Kewen Li, Chenxi Chen, Wenying Liu, Xianghua Fang, and Qinghua Lu. Software defect prediction using fuzzy integral fusion based on GA-FM. *Wuhan University Journal of Natural Sciences*, 19(5):405–408, 2014.

[98] Francesco Cavrini, Luigi Bianchi, Lucia Rita Quitadamo, and Giovanni Saggio. A Fuzzy Integral Ensemble Method in Visual P300 Brain-Computer Interface. *Computational Intelligence and Neuroscience*, 2016:1–9, dec 2016.

[99] Paweł Karczmarek, Witold Pedrycz, Marek Reformat, and Elaheh Akhoundi. A study in facial regions saliency: A fuzzy measure approach. *Soft Computing*, 18:379–391, 2014.

[100] Z.Y Wang and N.F Xiao. Fuzzy Integral-based Neural Network Ensemble for Facial Expression Recognition. *Proceedings of the International Conference on Computer Information Systems and Industrial Applications*, 2015.

[101] Anthony J. Pinar, Joseph Rice, Timothy C. Havens, Matthew Masarik, Joseph Burns, and Derek T. Anderson. Explosive hazard detection with feature and decision level fusion, multiple kernel learning, and fuzzy integrals. *IEEE Symposium Series on Computational Intelligence, SSCI 2016*, pages 1–8, dec 2016.

[102] American Cancer Society. Global Cancer Facts & Figures 2015. *Cancer Facts & Figures 2015*, (3rd edition), 2015.

[103] Worldwide cancer statistics Cancer Research UK. http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer. last visited: 15/02/2019.

[104] A R Green, D Soria, J Stephen, D G Powe, C C Nolan, I Kunkler, J Thomas, G R Kerr, W Jack, D Cameron, T Piper, G R Ball, J M Garibaldi, E A Rakha, J MS Bartlett, and I O Ellis. Nottingham Prognostic Index Plus: Validation of a clinical decision making tool in breast cancer in an independent series. *The Journal of Pathology: Clinical Research*, 2(1):32–40, 2016.

[105] Khairul A Rasmani, Jonathan M Garibaldi, Qiang Shen, and Ian O Ellis. Linguistic rulesets extracted from a quantifier-based fuzzy classification system. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*, pages 1204–1209. IEEE, 2009.

[106] E A Rakha, D Soria, A R Green, C Lemetre, D G Powe, C C Nolan, J M Garibaldi, G Ball, and I O Ellis. Nottingham Prognostic Index Plus (NPI+): a modern clinical decision making tool in breast cancer. *British journal of cancer*, 110(7):1688–97, 2014.

[107] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the AMIA Symposium. American Medical Informatics Association*, 95(25):14863–8, 1998.

[108] Meltzer Bittner, P Meltzer, Y Chen, Y Jiang, E Seftor, M Hendrix, M Radmacher, Rm Simon, Z Yakhini, A Ben-Dor, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536, 2000.

[109] C M Perou, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C a Rees, J R Pollack, D T Ross, H Johnsen, L a Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, a L Børresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[110] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530, 2002.

[111] T Sorlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P E Lønning, and A L Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–74, 2001.

[112] C Sotiriou, S Y Neo, L M McShane, E L Korn, P M Long, A Jazaeri, P Martiat, S B Fox, A L Harris, and E T Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–10398, 2003.

[113] S Monti, P Tamayo, J Mesirov, and T Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.

[114] P Kellam, X Liu, N Martin, C Orengo, S Swift, and A Tucker. Comparing, Contrasting and Combining Clusters in Viral Gene Expression. *Proceedings of the Sixth Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, oct 2001.

[115] Daniele Soria and Jonathan M. Garibaldi. A novel framework to elucidate core classes in a dataset. *IEEE World Congress on Computational Intelligence(WCCI) - Congress on Evolutionary Computation (CEC)*, 2010.

[116] Khairul A Rasmani and Qiang Shen. Modifying weighted fuzzy subsethood-based rule models with fuzzy quantifiers. In *Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on*, volume 3, pages 1679–1684. IEEE, 2004.

[117] Jonathan M Garibaldi, Daniele Soria, and Khairul A Rasmani. Consensus clustering and fuzzy classification for breast cancer prognosis. 2010.

[118] van Huysduynen et al. Measuring driving styles: A validation of the multidimensional driving style inventory. In *Proc of the 7th Int Conf on Auto User Interf and Interactive Vehicular Apps*, pages 257–264, 2015.

[119] Fridulv Sagberg, Selpi, Giulio Francesco Bianchi Piccinini, and Johan Engström. A review of research on driving styles and road safety. *Human Factors*, 57(7):1248–1275, 2015.

[120] R. Kalsoom and Z. Halim. Clustering the driving features based on data streams. In *Multi Topic Conference (INMIC), 2013 16th International*, pages 89–94, Dec 2013.

[121] German Castignani, Thierry Derrmann, Raphaël Frank, and Thomas Engel. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. *IEEE Int Transp Sys Magazine*, 7(1):91–102, 2015.

[122] Zahid Halim, Rizwana Kalsoom, and Abdul Rauf Baig. Profiling drivers based on driver dependent vehicle driving features. *Appl Int*, 44(3):645–664, 2016.

[123] Adrian B Ellison, SP Greaves, and Rhonda Daniels. Profiling drivers' risky behaviour towards all road users. In *A safe system: expanding the reach: Australasian College of Road Safety national conference*, 2012.

[124] C. Saiprasert, S. Thajchayapong, T. Pholprasit, and C. Tanprasert. Driver behaviour profiling using smartphone sensory data in a v2i environment. In *2014 Int Conf on Connected Vehicles and Expo (ICCVE)*, pages 552–557, Nov 2014.

[125] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[126] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[127] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.

[128] Shaoxin Li, Junliang Xing, Zhiheng Niu, Shiguang Shan, and Shuicheng Yan. Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 222–230, 2015.

[129] Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 158–165. IEEE, 2018.

[130] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[131] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[132] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017.

[133] Shervin Rahimzadeh Arashloo and Josef Kittler. Energy normalization for pose-invariant face recognition based on mrf model image matching. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1274–1280, 2011.

[134] Hossein Tahani and James M Keller. Information fusion in computer vision using the fuzzy integral. *IEEE Transactions on systems, Man, and Cybernetics*, 20(3):733–741, 1990.

[135] Paweł Karczmarek, Adam Kiersztyn, and Witold Pedrycz. An evaluation of fuzzy measure for face recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10245 LNAI:668–676, jun 2017.

[136] Paweł Karczmarek, Adam Kiersztyn, and Witold Pedrycz. Generalized Choquet Integral for Face Recognition. *International Journal of Fuzzy Systems*, (July):1–9, aug 2017.

[137] Yiliang Chen, Zichang Tan, Alex Po Leung, Jun Wan, Jianguo Zhang, et al. Multi-region ensemble convolutional neural networks for high accuracy age estimation. 2017.

[138] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Multi-region ensemble convolutional neural network for facial expression recognition. *arXiv preprint arXiv:1807.10575*, 2018.

[139] Dechang Chen, Kai Xing, Donald Henson, Li Sheng, Arnold M Schwartz, and Xiuzhen Cheng. Developing prognostic systems of cancer patients by ensemble clustering. *Journal of biomedicine & biotechnology*, june 2009.

[140] Darina Vuong, Peter T. Simpson, Benjamin Green, Margaret C. Cummings, and Sunil R. Lakhani. Molecular classification of breast cancer. *Virchows Archiv*, 465(1):1–14, jul 2014.

[141] S E. Elsheikh, A R. Green, E A. Rakha, D G. Powe, R A. Ahmed, H M. Collins, D Soria, J M. Garibaldi, C E. Paish, A A. Ammar, M J. Grainge, G R. Ball, M K. Abdelghany, L Martinez-Pomares, D M. Heery, and I O. Ellis. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. *Cancer Research*, 69(9):3802–3809, 2009.

[142] S. Gopal Krishna Patro and Kishore Kumar Sahu. Normalization: A Preprocessing Stage. *arXiv preprint arXiv:1503.06462.*, page 4, 2015.

[143] Stephen Swift, Allan Tucker, Veronica Vinciotti, Nigel Martin, Christine Orengo, Xiaohui Liu, and Paul Kellam. Consensus clustering and functional interpretation of gene-expression data. *Genome biology*, 5(11):R94, 2004.

[144] Zhao Yang and Ming Zhou. Kappa statistic for clustered matched-pair data. *Statistics in Medicine*, 33(15):2612–2633, jul 2014.

[145] Martin Krzywinski and Naomi Altman. Visualizing samples with box plots. *Nature Publishing Group*, 11(2), 2014.

[146] Thomas Sellke, M. J Bayarri, and James O Berger. Values for Testing Precise Null Hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[147] Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi, and Elvia M. Quiroz. Internal versus External cluster validation indexes. *International Journal of Computers and Communications*, 5(1):27–34, 2011.

[148] Dongkuan Xu and Yingjie Tian. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2:165–193, jun 2015.

[149] Diego Raphael Amancio, Cesar Henrique Comin, Dalcimar Casanova, Gonzalo Travieso, Odemir Martinez Bruno, Francisco Aparecido Rodrigues, and Luciano Da Fontoura Costa. A systematic comparison of supervised classifiers. *PLoS ONE*, 9(4):e94137, apr 2014.

[150] M. Lichman. {UCI} Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2013.

[151] Department of Transport. Road freight statistics. *Available at: https://www.gov.uk/government/collections/road-freight-domestic-and-international-statistics*, Last accessed 30 Sep 2016.

[152] Isaaac Triguero, Grazziela Figueredo, Mohammad Mesgarpour, Jonathan M. Garibaldi, and Robert I. John. Vehicle incident hot spots identification: An approach for big data. In *The 11th IEEE Int Conf On Big Data Sci and Eng*, 2017.

[153] Grazziela P Figueredo et al. Detecting danger in roads: An immune-inspired technique to identify heavy good vehicles incident hot spots. *IEEE Trans on Emmerging Topics in Comp Int*, 2017.

[154] Microlise. Microlise (telematics, transport and fleet management solutions). *Available at: http://www.microlise.com/*, Last accessed 27 Sep 2016.

[155] Chien-Ming Tseng et al. A comprehensive analysis of factors leading to speeding offenses among large-truck drivers. *Trans Research Part F: Traffic Psychology and Behaviour*, 38:171 – 181, 2016.

[156] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[157] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. Efficient knn classification algorithm for big data. *Neurocomputing*, 195:143–148, 2016.

[158] Kevin Gurney. *An introduction to neural networks*. CRC press, 2014.

[159] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[160] Timothy C Havens, Derek T Anderson, and Christian Wagner. Data-informed fuzzy measures for fuzzy integration of intervals and fuzzy numbers. *IEEE Transactions on Fuzzy Systems*, 23(5):1861–1875, 2015.

[161] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[162] Michael L Matthews. Discrimination of identikit constructions of faces: Evidence for a dual processing strategy. *Perception & Psychophysics*, 23(2):153–161, 1978.

[163] Yu Chien Ko, Hamido Fujita, and Gwo Hshiung Tzeng. An extended fuzzy measure on competitiveness correlation based on WCY 2011. *Knowledge-Based Systems*, 37:86–93, 2013.

[164] Derek T. Anderson, James M. Keller, and Timothy C. Havens. Learning fuzzy-valued fuzzy measures for the fuzzy-valued sugeno fuzzy integral. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6178 LNAI:502–511, 2010.

[165] IO Ellis, M Galea, N Broughton, A Locker, RW Blamey, and CW Elston. Pathological prognostic factors in breast cancer. ii. histological type. relationship with survival in a large study with long-term follow-up. *Histopathology*, 20(6):479–489, 1992.

[166] Therese Sorlie, Robert Tibshirani, Joel Parker, Trevor Hastie, JS Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8418–8423, 2003.

[167] Jason I Herschkowitz, Karl Simin, Victor J Weigman, Igor Mikaelian, Jerry Usary, Zhiyuan Hu, Karen E Rasmussen, Laundette P Jones, Shahin Assefnia, Subhashini Chandrasekharan, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome biology*, 8(5):R76, 2007.

[168] Maggie CU Cheang, Stephen K Chia, David Voduc, Dongxia Gao, Samuel Leung, Jacqueline Snider, Mark Watson, Sherri Davies, Philip S Bernard, Joel S Parker, et al. Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, 101(10):736–750, 2009.

[169] Jack Cuzick, Mitch Dowsett, Silvia Pineda, Christopher Wale, Janine Salter, Emma Quinn, Lila Zabaglo, Elizabeth Mallon, Andrew R Green, Ian O Ellis, et al. Prognostic value of a combined estrogen receptor, progesterone receptor, ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer. *J Clin Oncol*, 29(32):4273–4278, 2011.

[170] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

[171] Peng Shi, Mingming Wang, Qinghui Zhang, and Jingzhong Sun. Lipid-rich carcinoma of the breast. a clinicopathological study of 49 cases. *Tumori Journal*, 94(3):342–346, 2008.

[172] M-C Mathieu, R Rouzier, A Llombart-Cussac, L Sideris, S Koscielny, JP Travagli, G Contesso, S Delaloge, and M Spielmann. The poor responsiveness of infiltrating lobular breast carcinomas to neoadjuvant chemotherapy can be explained by their biological profile. *European journal of cancer*, 40(3):342–351, 2004.

[173] Ola H Negm, Abir A Muftah, Mohammed A Aleskandarany, Mohamed R Hamed, Dena AJ Ahmad, Christopher C Nolan, Maria Diez-Rodriguez, Patrick J Tighe, Ian O Ellis, Emad A Rakha, et al. Clinical utility of reverse phase protein array for molecular classification of breast cancer. *Breast cancer research and treatment*, 155(1):25–35, 2016.

[174] Heiko A Mannsperger, Stephan Gade, Frauke Henjes, Tim Beissbarth, and Ulrike Korf. Rppanalyzer: Analysis of reverse-phase protein array data. *Bioinformatics*, 26(17):2202–2203, 2010.

Appendix A

Ki67 importance in Breast Cancer Stratification

A.1 Introduction

Determination of breast cancer (BC) patients with high risk of recurrence, metastasis and therapy resistance can avoid potential over-treatment and improve their survival. The main step toward personalised medicine is the identification of breast cancer biological diversity and proper categorisation of patients. Current BC classification is mainly based on clinico-pathological criteria which are mainly tumour size, lymph node involvement, histological type and grade [165, 106, 104]. In addition, routine immunohistochemistry is used for identification of predictive markers: Oestrogen Receptor (ER), Progesterone Receptor (PR) and HER2 receptors dividing patients into therapy groups accordingly. Other markers have also been evaluated, such as Ki67, the use of which has been figured out as a prognostic or predictive biomarker in invasive breast cancer. Furthermore, introduction of molecular subtypes has opened a new avenue for clinicians to refine breast cancer classification and develop subgroups with the aim of individualised therapy. Perou et.al classified breast cancer into intrinsic subtypes (luminal, basal-like, and HER2-positive) based on gene expression profiling and are associated with patient outcome [109]. They mentioned that within the 1,753 gene cluster, the largest distinctive cluster of genes was the proliferation cluster which its expression correlates with cellular proliferation rates and well correlated with the mitotic index. This cluster included Ki67 which is used as immunohistochemical markers of cell proliferation.

Subsequently, further divisions have been added: luminal A, luminal B and claudin-low subtypes [111, 166, 167]. Luminal B tumours have a poorer prognosis and higher proliferation than luminal A tumours. Cheang and colleagues developed a clinically practical immunohistochemistry assay to differentiate between luminal A and B breast cancers using Ki67 [168]. They concluded that the expression of ER, PR, HER2, and Ki67 proteins appear to discriminate luminal A from luminal B breast cancer subtypes. An immunohistochemistry test, HC4, of these four biomarkers has been developed by Cuzick et. al. [169]. The result provided by the IHC4 score were found to be similar to that presented by Oncotype DX which is a commercially available multigene assay of 21 genes including Ki67 as a primary gene in the cell proliferation group which are heavily weighted in the formula used to estimate the tumour recurrence score [170]. Furthermore, Ki67 expression is used as a descriptive parameter in rare epithelial subtypes of the WHO classification of breast carcinoma. In lipid-rich subtype, Li67 is over expressed while in lobular carcinoma is low [171, 172].

The aim of this study is to use the same markers for clustering using RPPA technique and to evaluate the importance (influence) of the biomarker '*Ki67*' on the classification results. For this, the pipeline proposed in Chapter4 is employed in three stages; first by considering 10 original protein biomarkers, second first ten plus Ki67 and third by considering all the 14 biomarkers. The difference in performance can indicate the answer to the aforementioned question.

A.2 Background

A.2.1 Reverse Phase Protein Microarray

Reverse Phase Protein Microarray (RPPA) was performed as described previously [173]. In brief, protein lysates were robotically spotted in triplicate onto nitrocellulose-coated glass slides with a microarraying robot (MicroGrid 610, Digilab Marlborough, MA, USA). Blocking of slides was performed using Super G blocking buffer (Grace Bio-labs, USA) overnight. Next, slides were incubated with the primary antibodies overnight at 4 °C with agitation at 20 rpm. Detection of the reaction was performed with infrared secondary antibodies (680 CW anti-mouse or 800 CW anti-rabbit antibody) diluted 1:5000 in washing buffer for 30 min at room temperature and visualised using an Odyssey high-resolution scanner (LI-COR, Lincoln, USA) at 21 lm resolution using both channels 700 (red) and 800 (green). Axon Genepix Pro-6 Microarray Image Analysis software was used to determine the average fluorescence pixel intensities of all spots on each array. Protein signals were finally

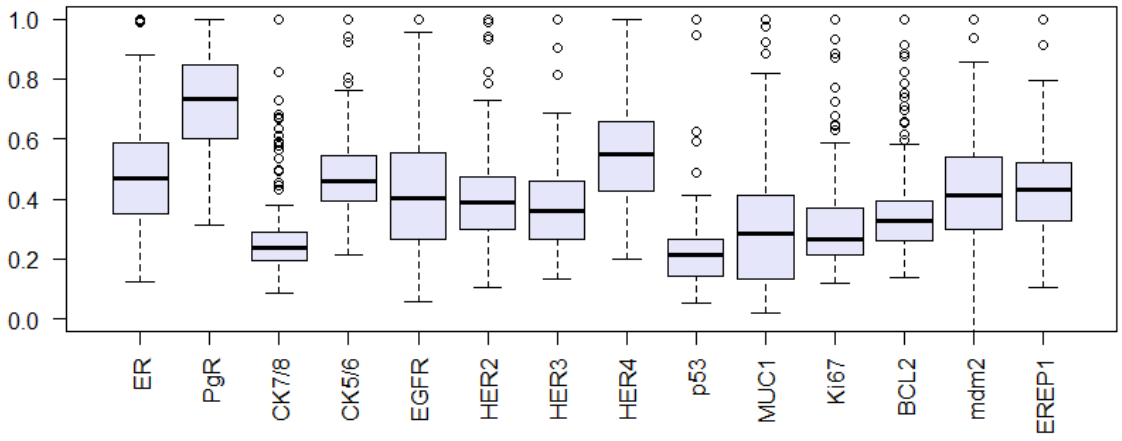


Figure A.1 Boxplot of the RPPA data

determined with background subtraction and normalisation to the internal housekeeping targets using RPPanalyzer [174].

A.2.2 Patient Samples

In this study, 143 cases of primary invasive BC prospectively collected from patients treated at Nottingham City Hospital (Nottingham prospective series) were used. The patients were under 70 years old and they received the same treatment according to the specific protocol which approved locally. Samples were provided by the Nottingham Health Science Biobank and the study was approved by the North West 7 Research Ethics Committee—Great Manchester Central (Nottingham Research Biorepository (NRB)—REC Reference number 10/H1008/72). RPPA differs from Immunohistochemistry in the method of scoring biomarkers, the former by computer while the latter being manual. The RPPA dataset consists of 14 protein biomarkers, first ten being the same as used in previous breast cancer datasets, 11th being Ki67, and three additional biomarkers namely BCL2, mdm2 and EREP1.

A boxplot of the range of biomarkers is shown in Figure A.1.

A.3 Data pre-processing

Before the analysis all the biomarkers were normalised on a scale of 0-1. There are number of ways to normalise including z-score, min-max and dividing by maximum. For the sake of simplicity we chose dividing by maximum normalisation. For each biomarker we first determine the maximum value throughout the dataset. For every patient divide each biomarker

by its corresponding maximum value. The data had no missing values, and contained three scores for each patient and each biomarker. The average of all the three values is computed to create the final dataset to be used for analysis.

A.4 Methodology

Steps followed for the analysis are as follows:

1. Two clustering algorithms: Partitioning Around Medoids (PAM) and K-means with centres from pruning of Hierarchical cluster tree are run on the dataset.
2. A consensus of both the clustering algorithms is performed (clusterfusion).
3. Clusterfusion is performed passing three cluster centres. Similar steps are repeated for creating eight clusters.
4. The results and discussion are reported for ten original biomarkers, original ten plus Ki67 and all the 14 biomarkers.
5. The next step is applying classification algorithms to the non-classified patients, and preform the consensus of the results.

A.5 Results

This section discusses the analysis on the Reverse Phase Protein Array (RPPA) data of 143 breast cancer patients.

A.5.1 Analysis on 10 biomarkers

The analysis on the ten original biomarkers are discussed in this section. The first step is to run the K-means and PAM clustering algorithms with three cluster centres of the dataset. A consensus among 99 patients is agreed (clusterfusion), whose distribution in classes shown in table A.1.

To distribute patients among seven classes, similar to the Nottingham case, the clustering algorithms are run with eight cluster centres. The first reason to choose eight over seven is to achieve similarity with Nottingham case: seven breast cancer classes and one NC class. The second reason being no consensus for one group upon passing seven cluster centres, leading

Method	Class 1	Class 2	Class 3
K-Means	51	46	46
PAM	58	32	53
Consensus	35	23	41

Table A.1 Results after division in three classes using 10 biomarkers

to six cancer groups. Upon passing eight centres, the pattern is repeated and no consensus is obtained among one group therefore forming seven classes and a NC one. Passing the eight centres to K-means and PAM, a consensus among 92 patients is agreed (clusterfusion). Table A.2 shows the distribution of patients among the seven (and NC) classes.

A comparison among the distribution of patients in three and seven classes is performed to compare the similarities among them. Classes 1, 3 and 5 of the seven class classification (scc) show similar distributions to class 2 of the three class classification (tcc). All the patients in classes 6 and 7 of the scc show similar distributions to class 3 of the tcc. A similar result is obtained for class 4 of the scc, which shows similar distributions to class 1 of the tcc. Class 2 of the scc showed a mixed distribution towards all the three classes of the tcc.

Among scc, classes 6 and 7 differ from each other in value of the biomarkers CK7/8, CK5/6 and EGFR. They are on the higher side in class 6 and on the lower side in class 7. The major difference among classes 1, 3 and 5 are the high values of PgR for class 1. Among classes 3 and 5, class 3 has low value of ER and high values of MUC1, and vice-versa for both in class 5 i.e. high values of ER and low values of MUC1. Classes 2 and 4 have very similar distribution, except for MUC1 biomarker: class 2 has higher values than class 4.

Method	NC	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
K-Means	14	24	25	6	28	6	33	7
PAM	11	21	28	6	23	17	25	12
Consensus	-	11	19	6	19	5	25	7

Table A.2 Results after division in seven (and NC) classes using 10 biomarkers

Two classification algorithms (ANN and Nearest Neighbour) are run on the distribution obtained after clusterfusion. A consensus of the results of the classification algorithms is performed (classifusion). After performing this step 32 more patients (in addition to 92 obtained after clusterfusion) are classified to one of the seven classes increasing the number to 124 patients. The boxplots of these additional patients for each class is similar to the ones obtained after clusterfusion. Total column in Table A.3 shows the overall distribution, classifying 124 patients in one of the seven classes.

Class	Clusterfusion	Classifusion	Total
1	11	11	22
2	19	7	26
3	6	1	7
4	19	3	22
5	5	4	9
6	25	6	31
7	7	0	7
NC	-	-	19

Table A.3 Results after combining with classifusion using 10 biomarkers

A.5.2 Analysis on 11 biomarkers

The analysis on 11 (ten original + Ki67) is discussed in this section. K-means and PAM algorithms are run on the dataset with three cluster centres. A consensus among 122 patients is agreed (clusterfusion), whose class distribution is shown in Table A.4.

Method	Class 1	Class 2	Class 3
K-Means	63	18	62
PAM	52	31	60
Consensus	52	18	52

Table A.4 Results after division in three classes using 11 biomarkers

To perform clusterfusion with eight cluster centres, K-means and PAM clustering algorithms are run on the dataset. A consensus among 99 patients is agreed, whose distribution in classes is shown in Table A.5. All the patients in classes 1 and 5 of the scc show similar distributions to class 1 of the tcc. Similarly, class 7 of the scc show similar distributions to class 2 of the tcc. Also, classes 3, 4 and 6 of the scc show similar distributions to all the patients in class 3 of the tcc. Class 2 of the scc show a mixed distribution towards all the three classes.

Among scc, classes 1 and 5 differ from each other in value of biomarkers PgR and EGFR. These are on the higher side in class 1 and on the lower side in class 5. The major difference among classes 3, 4 and 6 are the high values of MUC1 for class 4. Class 3 has high values of ER and HER2. Classes 2 and 7 have dissimilar distributions. The next step is to implement classifusion, which classifies additional 32 patients in one of the seven classes. Total column in Table A.6 shows the overall distribution.

Method	NC	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
K-Means	4	33	24	12	11	24	22	13
PAM	19	25	23	10	11	22	21	12
Consensus	-	25	16	7	6	18	16	11

Table A.5 Results after division in seven (and NC) classes using 11 biomarkers

Class	Clusterfusion	Classifusion	Total
1	25	7	32
2	16	5	21
3	7	2	9
4	6	6	12
5	18	4	22
6	16	6	22
7	11	2	13
NC	-	-	12

Table A.6 Results after combining with classifusion using 11 biomarkers

A.5.3 Analysis on 14 biomarkers

The analysis on 14 biomarkers is discussed in this section to see the affect of other biomarkers on the classification. K-means and PAM are run on the dataset with three cluster centres. A consensus among 123 patients is agreed (clusterfusion), whose distribution in classes is shown in Table A.7.

Method	Class 1	Class 2	Class 3
K-Means	67	20	56
PAM	53	29	61
Consensus	53	20	50

Table A.7 Results after division in three classes using 14 biomarkers

Similar to analysis on 10 and 11 biomarkers, clusterfusion is performed with eight cluster centres. K-means and PAM clustering algorithms have consensus among 94 patients, whose distribution in classes is shown in Table A.8. All the patients in classes 6 and 7 of the scc show similar distributions to class 2 of the tcc. Similarly, all the patients in classes 2 and 3 of the scc show similar distributions to class 3 of the tcc. Also, all the class 4 patients of the scc show similar distributions to class 1 of the tcc. Class 1 and 5 of the scc show a mixed distribution towards classes 1 and 3 of the tcc.

In scc, classes 6 and 7 differ from each other in value of biomarkers ER, HER2, HER4 and EREP1. These are on the higher side in class 7 and on the lower side in class 6. The major difference among the classes 2 and 3 are high values of ER and low value of MUC1 for class 2 and vice-versa for class 3. Among classes 1 and 5, class 1 has high value of EGFR, but on the lower side for the class 5. Class 4 has similar distribution to classes 1 and 5, except for the PgR biomarker which are on higher side in class 4. The next step is to implement classifusion, which classifies additional 31 patients in one of the seven classes. Total column in Table A.9 shows the overall distribution.

Method	NC	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
K-Means	4	24	19	7	31	37	7	14
PAM	22	25	18	13	27	23	9	6
Consensus	-	15	15	6	26	19	7	6

Table A.8 Results after division in seven (and NC) classes using 14 biomarkers

Class	Clusterfusion	Classifusion	Total
1	15	6	21
2	15	8	23
3	6	2	8
4	26	7	33
5	19	8	27
6	7	0	7
7	6	0	6
NC	-	-	18

Table A.9 Results after combining with classifusion using 14 biomarkers

A.5.4 Discussion

For division of patients in three classes, clusterfusion gave best results with 14 and 11 biomarkers case. 11 biomarkers classified 122 patients, while 14 biomarkers classified one extra i.e. 123. Worst performance was obtained with 10 biomarkers, which classified 99 patients. For division of patients in seven classes (and NC), 11 biomarkers gave best result after clusterfusion, classifying 99 patients. It gave best results after classifusion as well, classifying a total of 131 patients. 10 and 14 biomarkers gave similar results after clusterfusion and classifusion.

Clinical Interpretation

For classification of patients in groups cluster validity indices suggest best number to be three, which is evident in the results. Analysis is done for both three and seven classes, but tcc results after clusterfusion outperforms scc in all the three cases. To check the importance of Ki67 biomarker in classification, analysis is done with and without Ki67. 11 biomarkers case outperformed 10 biomarkers in both three and seven class division. Thus, it can be inferred that addition of Ki67 biomarkers improves the classification results.