Enhancing Fuzzy Associative Rule Mining Approaches for Improving Prediction Accuracy

B I Sowan

PhD

2011

Enhancing Fuzzy Associative Rule Mining Approaches for Improving Prediction Accuracy

Integration of Fuzzy Clustering, Apriori and Multiple Support Approaches to Develop an Associative Classification Rule Base

Bilal Ibrahim Sowan

BSc, MSc

A thesis submitted for the degree of

Doctor of Philosophy

School of Computing, Informatics & Media

University of Bradford

2011

July 2011

University of Bradford

Dedication

To my father

To the memory of my mother

To my brothers and sisters

With much love, affection and appreciation.

Abstract

Building an accurate and reliable model for prediction for different application domains, is one of the most significant challenges in knowledge discovery and data mining. This thesis focuses on building and enhancing a generic predictive model for estimating a future value by extracting association rules (knowledge) from a quantitative database. This model is applied to several data sets obtained from different benchmark problems, and the results are evaluated through extensive experimental tests.

The thesis presents an incremental development process for the prediction model with three stages. Firstly, a Knowledge Discovery (KD) model is proposed by integrating Fuzzy C-Means (FCM) with Apriori approach to extract Fuzzy Association Rules (FARs) from a database for building a Knowledge Base (KB) to predict a future value. The KD model has been tested with two road-traffic data sets.

Secondly, the initial model has been further developed by including a diversification method in order to improve a reliable FARs to find out the best and representative rules. The resulting Diverse Fuzzy Rule Base (DFRB) maintains high quality and diverse FARs offering a more reliable and generic model. The model uses FCM to transform quantitative data into fuzzy ones, while a Multiple Support Apriori (MSapriori) algorithm is adapted to extract the FARs from fuzzy data. The correlation values for these FARs are calculated, and an efficient orientation for filtering FARs is performed as a post-processing method. The FARs diversity is maintained through the clustering of FARs, based on the concept of the sharing function technique used in multi-objectives optimization. The best and the most diverse FARs are obtained as the DFRB to utilise within the Fuzzy Inference System (FIS) for prediction.

The third stage of development proposes a hybrid prediction model called Fuzzy Associative Classification Rule Mining (FACRM) model. This model integrates the

improved Gustafson-Kessel (G-K) algorithm, the proposed Fuzzy Associative Classification Rules (FACR) algorithm and the proposed diversification method. The improved G-K algorithm transforms quantitative data into fuzzy data, while the FACR generate significant rules (Fuzzy Classification Association Rules (FCARs)) by employing the improved multiple support threshold, associative classification and vertical scanning format approaches. These FCARs are then filtered by calculating the correlation value and the distance between them. The advantage of the proposed FACRM model is to build a generalized prediction model, able to deal with different application domains. The validation of the FACRM model is conducted using different benchmark data sets from the University of California, Irvine (UCI) of machine learning and KEEL (Knowledge Extraction based on Evolutionary Learning) repositories, and the results of the proposed FACRM are also compared with other existing prediction models. The experimental results show that the error rate and generalization performance of the proposed model is better in the majority of data sets with respect to the commonly used models.

A new method for feature selection entitled Weighting Feature Selection (WFS) is also proposed. The WFS method aims to improve the performance of FACRM model. The prediction performance is improved by minimizing the prediction error and reducing the number of generated rules. The prediction results of FACRM by employing WFS have been compared with that of FACRM and Stepwise Regression (SR) models for different data sets. The performance analysis and comparative study show that the proposed prediction model provides an effective approach that can be used within a decision support system.

Acknowledgment

First and foremost, I would like to thank GOD (ALLAH) the most merciful, the most generous, the most gracious, and the most compassionate for providing me the blessings, courage, strength and help to complete this work.

I would like to take this opportunity to express my sincere gratitude, gratefulness and appreciation to my supervisors, Dr. Keshav Dahal and Prof. Alamgir Hossain for their valuable advice, patience, generosity with their time and effort, kindness, expert guidance and useful feedback throughout my PhD journey. Our regular meetings and useful discussions provided me with encouragement, constant support and an integral view of research helping me pursue the right direction in. In fact, without their encouragement and great support, I would not have been able to achieve this work.

I would like to give special and warm appreciation to my father for his constant encouragement and moral support. His spiritual support and prayers are the main motivation for my success in my life and studies. I am indebted to my brothers and sisters for their loving support and inspiration.

I would like to sincerely thank all my friends and colleagues at the University of Bradford who have provided a lot of collaboration and emotional support.

Finally, it is my pleasure to acknowledge the financial support for this research from Applied Science University (ASU) of Jordan.

Table of Contents

ABSTRACT .... **I**

ACKNOWLEDGMENT .... **III**

TABLE OF CONTENTS .... **IV**

LIST OF FIGURES .... **VIII**

LIST OF TABLES .... **XI**

LIST OF ABBREVIATIONS .... **XIV**

CHAPTER ONE **1**

1 INTRODUCTION **1**

1.1 BACKGROUND TO THE RESEARCH PROBLEM 1

1.2 RESEARCH MOTIVATION AND JUSTIFICATION 3

1.3 RESEARCH AIMS AND OBJECTIVES 7

1.4 RESEARCH CONTRIBUTIONS 9

1.5 RESEARCH PROCESS 10

1.6 STATEMENT OF ETHICS 13

1.7 THESIS ORGANIZATION 14

1.8 PUBLICATIONS 15

CHAPTER TWO **17**

2 REVIEW OF EXISTING TECHNIQUES AND EVALUATION CRITERIA **17**

2.1 CHAPTER OVERVIEW 17

2.2 KNOWLEDGE DISCOVERY AND DATA MINING 18

2.2.1 Data Mining Tasks *19*

2.3 ASSOCIATION RULE MINING 22

2.3.1 Apriori Algorithm *25*

2.3.2 AprioriTid Algorithm *27*

2.3.3 Frequent Patterns Growth (FP-Growth) *29*

2.3.4 Multiple Support Approaches *30*

2.4 FUZZY LOGIC (FL) 35

2.5 CLUSTERING TECHNIQUES 36

2.5.1 Fuzzy C-Means (FCM) Algorithm *37*

2.5.2 The Gustafson-Kessel (G-K) Algorithm *39*

2.6 FUZZY INFERENCE AND EXPERT SYSTEM 39

2.7 FUZZY ASSOCIATION RULES (FARS) 41

2.8 SIMILARITY 43

2.9 ASSOCIATIVE CLASSIFICATION APPROACH 45

2.10 VERTICAL DATA REPRESENTATION 47

2.11 ARTIFICIAL NEURAL NETWORK 49

2.12 SUPPORT VECTOR MACHINE 51

2.13 STEPWISE REGRESSION 52

2.14 CLASSIFICATION AND REGRESSION TREES 52

2.15 FEATURE SELECTION METHODS 53

2.16 EVALUATION CRITERIA 54

2.17 SUMMARY AND CONCLUSIONS 57

CHAPTER THREE 59

3 PREDICTION MODELS BASED ON FUZZY ASSOCIATION RULES MINING 59

3.1 CHAPTER OVERVIEW 59

3.2 INTRODUCTION AND RELATED WORK 60

3.3 CASE STUDIES 63

3.3.1 Road Traffic Case Studies *63*

3.3.2 Abalone Benchmark Data *66*

3.4 THE PROPOSED FIRST PREDICTION MODEL 66

3.4.1 The Model Description *66*

3.4.2 Experimental Results and Analysis *70*

3.5 THE PROPOSED SECOND PREDICTION MODEL 78

3.5.1 The Model Description *78*

3.5.2 Experimental Results and Comparative Study *86*

3.6 SUMMARY AND CONCLUSION 91

CHAPTER FOUR 93

4 PREDICTION MODEL WITH IMPROVED MULTIPLE SUPPORT ASSOCIATIVE CLASSIFICATION APPROACH 93

4.1 CHAPTER OVERVIEW 93

4.2 INTRODUCTION 94

4.3 RELATED LITERATURE WORK 96

4.4 THE PROPOSED FUZZY ASSOCIATIVE CLASSIFICATION RULES MINING (FACRM) MODEL 99

4.4.1 The Model Description *99*

4.4.2 Experimental Results and Comparative Study *108*

4.5 SUMMARY AND CONCLUSIONS 121

CHAPTER FIVE 123

5 FEATURE SELECTION METHOD FOR ENHANCING THE PREDICTION MODEL 123

5.1 CHAPTER OVERVIEW 123

5.2 INTRODUCTION 123

5.3 LITERATURE REVIEW AND RELATED WORK 125

5.4 THE PROPOSED METHODOLOGY 130

5.5 EXPERIMENTAL RESULTS AND ANALYSIS 134

5.6 SUMMARY AND CONCLUSION 140

CHAPTER SIX 142

6 CONCLUSIONS AND FUTURE WORK 142

6.1 OBSERVATIONS AND CONCLUSIONS 142

6.2 FUTURE WORK 148

REFERENCES 152

List of Figures

Figure 2.1 Generic steps of the KDD. 18

Figure 2.2 Data mining tasks. 20

Figure 2.3 Apriori algorithm (Agrawal and Srikant, 1994). 26

Figure 2.4 Apriori algorithm example. 27

Figure 2.5 AprioriTid algorithm example. 28

Figure 2.6 FP-Growth example. 30

Figure 2.7 An example of fuzzy sets. 36

Figure 2.8 Two clusters of the data values from Table 2-5. 38

Figure 2.9 FIS diagram. 40

Figure 2.10 An example of multiple-level idea. 43

Figure 2.11 Fuzzy sets. 45

Figure 2.12 Vertical data format. 48

Figure 2.13 The structure of Multi-Layer Perception (MLP) ANN. 50

Figure 2.14The classification in SVM. 51

Figure 3.1 Information of the input road traffic data set. 64

Figure 3.2 Analysis of the road traffic data set for Demand 1 and Demand 2. 65

Figure 3.3 Analysis of the road traffic data set for Density 1, Density 2, and ATT. 65

Figure 3.4 The proposed first KD model steps. 67

Figure 3.5 The proposed first KD model. 69

Figure 3.6 The membership functions for each field used in this case study. (a) the fuzzy sets of Demand 1. (b) the fuzzy sets of Demand 2. (c) the fuzzy

SETS OF DENSITY 1. (D) THE FUZZY SETS OF DENISTY 2. (E) THE FUZZY SETS AVERAGE TOTAL TIME. 72
FIGURE 3.7 THE DIFFERENCE BETWEEN REAL AND PREDICTED ATT. 75
FIGURE 3.8 THE MAPE OVER DIFFERENT *MINSUPP* AND *MINCONF*. 76
FIGURE 3.9 THE MAPE OVER DIFFERENT *MINCONF* AND *MINSUPP*. 76
FIGURE 3.10 THE MAPE OVER DIFFERENT *MINSUPP* AND *MINCONF*. 77
FIGURE 3.11 THE MAPE OVER DIFFERENT *MINSUPP* AND *MINCONF* OF BOTH SMALL AND LARGE DATA SET. 78
FIGURE 3.12 THE PROPOSED SECOND PREDICTION MODEL WITH DFRB APPROACH PHASES. 81
FIGURE 3.13 EXTRACTION OF ROBUST AND DIVERSE FUZZY RULE BASE ALGORITHM. 85
FIGURE 3.14 SELECTION OF THE FUZZY RULE BASE WITH ROBUSTNESS AND DIVERSITY ... 86
FIGURE 3.15 THE SENSITIVITY OF MAPE AND *MINCORR*. 89
FIGURE 4.1 THE PROPOSED FACRM MODEL FOR PREDICTION. 101
FIGURE 4.2 FACR ALGORITHM. 106
FIGURE 4.3 THE COMPARISON BETWEEN FACRM AND ANN USING MAPE VALUES FOR ALL DATA SETS. 110
FIGURE 4.4 THE COMPARISON BETWEEN FACRM AND SVM USING MAPE VALUES FOR ALL DATA SETS. 111
FIGURE 4.5 THE COMPARISON BETWEEN FACRM AND SR USING MAPE VALUES FOR ALL DATA SETS. 111
FIGURE 4.6 THE INPUT GAS FLOW RATE $\mathbf{x}(\mathbf{t})$. 118
FIGURE 4.7 THE OUTPUT OF CO2 CONCENTRATION $\mathbf{y}(\mathbf{t})$. 118
FIGURE 4.8 THE DIFFERENCE BETWEEN THE ACTUAL AND PREDICTED VALUES. 119
FIGURE 5.1 THE PROPOSED WFS METHOD. 131
FIGURE 5.2 THE DEFINITION OF THE PROPOSED WFS METHOD. 131

FIGURE 5.3 FLOW CHART OF WFS METHOD. 132

FIGURE 5.4 AN EXAMPLE OF WFS METHOD. 133

FIGURE 5.5 THE WFS METHOD. 134

FIGURE 5.6 THE COMPARISON BETWEEN FACRM-FULL SET, FACRM-WFS METHOD AND SR USING MAPE VALUES FOR ALL DATA SETS. 137

List of Tables

TABLE 2-1 TRANSACTION ITEMS. 24

TABLE 2-2 BOOLEAN TRANSACTION ITEMS. 24

TABLE 2-3 ASSOCIATION RULES.......... 24

TABLE 2-4 DIFFERENCE BETWEEN APRIRORI, MSAPRIORI AND IMSAPRIORI. 34

TABLE 2-5 FCM EXAMPLE. 38

TABLE 3-1 PART OF THE ROAD TRAFFIC DATA SET. 65

TABLE 3-2 STATISTICAL INFORMATION OF THE ROAD TRAFFIC DATA SET. 65

TABLE 3-3 CENTERS OF THE DATA SET WITH MINIMUM AND MAXIMUM VALUES FOR EACH FIELD. 71

TABLE 3-4 PART OF FUZZY DATA SET FOR THE ROAD TRAFFIC. 72

TABLE 3-5 PART OF $\mathbf{C1}$. 73

TABLE 3-6 PART OF $\mathbf{L1}$. 73

TABLE 3-7 SUPPORT VALUE CALCULATION. 73

TABLE 3-8 PART OF $\mathbf{C2}$. 73

TABLE 3-9 PART OF $\mathbf{L2}$. 73

TABLE 3-10 PART OF $\mathbf{C3}$. 74

TABLE 3-11 PART OF $\mathbf{L3}$. 74

TABLE 3-12 PART OF $\mathbf{CV}$CALCULATING. 74

TABLE 3-13 STATISTICAL GAUGES OF THE PREDICTION EVALUATION 78

TABLE 3-14 FUZZY DATA SET. 82

TABLE 3-15 THE SENSITIVITY OF MAPE (%) ADAPTING *MINCORR*. 88

TABLE 3-16 THE SENSITIVITY OF MAPE (%) ADAPTING *MINCORR* AND DIVERSE FARS. 88

TABLE 3-17 THE SENSITIVITY OF MAPE (%) AND Σ VALUE OF *MINCORR*= 0.6 WITH DIVERSE FARS. 88

Table 3-18 The sensitivity of MAPE (%) over different numbers of FARs and diverse FARs. 90
Table 3-19 Calculation of the evaluation criteria. 90
Table 3-20 Calculation of MAE and MAPE (%). 90
Table 4-1 Fuzzy Data (FD). 103
Table 4-2 Fuzzy Term Set 1 (FTS_1/$EFTS_1$). 103
Table 4-3 Fuzzy Rule Term Set 1 ($FRTS_1$/$FRTS_1$`). 103
Table 4-4 Fuzzy Term Set 2 (FTS_2/$EFTS_2$). 104
Table 4-5 Fuzzy Rule Term Set 2 ($FRTS_2$). 105
Table 4-6 Basic information of the data sets from UCI and Keel repositories. 109
Table 4-7 MAPE (%) values of prediction models for all data sets. 110
Table 4-8 MdAPE (%) values of prediction models for all data sets. 112
Table 4-9 RMSE values of prediction models for all data sets. 112
Table 4-10 MAE values of prediction models for all data. 113
Table 4-11 The data sets numbers that produced minimum error values out of 14 data sets. 113
Table 4-12 The average of mean error values of all prediction models. 114
Table 4-13 The number of data sets out of 14 data sets that FACRM beats other models. 114
Table 4-14 The comparison of FACRM and CART. 115
Table 4-15 Number of rules generated in FACRM and CART. 116
Table 4-16 Statistical significance using Wilcoxon rank sum test. 117
Table 4-17 The comparison of FACRM with different models. 120
Table 5-1 Full features, selected features in all data sets used in FACRM and SR. The values of LS and *minconf* are also used in FACRM for each data set. 135

TABLE 5-2 MAPE (%) VALUES OF PREDICTION MODELS FOR ALL DATA SETS USING FULL SET AND SELECTED FEATURES. 136

TABLE 5-3 MDAPE (%) VALUES OF PREDICTION MODELS IN ALL DATA USING FULL SET AND SELECTED FEATURES. 136

TABLE 5-4 RMSE VALUES OF PREDICTION MODELS FOR ALL DATA SETS USING FULL SET AND SELECTED FEATURES. 136

TABLE 5-5 MAE VALUES OF PREDICTION MODELS FOR ALL DATA SETS USING FULL SET AND SELECTED FEATURES. 137

TABLE 5-6 THE DISCOVERED AND DIVERSE RULES GENERATED BY FACRM USING FULL SET AND SELECTED FEATURES. 138

List of Abbreviations

AC	Associative Classification
AI	Artificial Intelligence
ANFIS	Adaptive Network based on Fuzzy Inference System
ANN	Artificial Neural Network
ATT	Average Travel Time
BF	Best First
CARS	Classification Association Rules
CART	Classification and Regression Tree
CBA	Classification Based on Associations
CFAR	Classification Fuzzy Association Rule
CFP-Growth	Conditional Frequent Pattern Growth
CFS	Correlation-based Feature Selection
CLARANS	Clustering LARge Applications based on RANdomized Search
CS	Coefficient Similarity
CSMC	Combination Strategy for Multi-class Classification
CV	Confidence Value
DFRB	Diverse Fuzzy Rule Base
DM	Data Mining
DR	Diverse Rules
DSS	Decision Support System
ES	Exhaustive Search
ECR	Equivalence Class Rule
EFDR	Enhanced Fuzzy Data Representation
EDP	Equi-Depth Partition
EDPFT	Equi-Depth Partition Fuzzy Terms algorithm

FACRM	Fuzzy Associative Classification Rule Mining
FACR	Fuzzy Associative Classification Rules
FARs	Fuzzy Association Rules
FCARs	Fuzzy Classification Association Rules
FCM	Fuzzy C-Means
FIS	Fuzzy Inference System
FL	Fuzzy Logic
FP-Growth	Frequent Pattern Growth
FP-Tree	Frequent Pattern Tree
FR	Fuzzy Rules
FRB	Fuzzy Rule Base
GA	Genetic Algorithm
G-K	Gustafson-Kessel
GS	Genetic Search
GS	Greedy Stepwise
ICFP-Growth	Improved Conditional Frequent Pattern Growth
IMSapriori	Improved Multiple Support apriori
KB	Knowledge Base
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
KEEL	Knowledge Extraction based on Evolutionary Learning
K-S	Kolmogorov-Smirnov
LIBSVM	LIBrary for Support Vector Machines
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCAR	Multi-class Classification based on Association Rules
MCDA	Multi-Criteria Decision Aid

MdAPE	Median Absolute Percentage Error
MFTS	Minimum Fuzzy Term Support
MI	Mutual Information
minconf	minimum confidence threshold
minCorr	minimum Correlation threshold
minsupp	minimum support threshold
MIS	Minimum Item Support
MLP	Multi-Layer Perception
mRMR	minimum Redundancy Maximum Relevance
MSapriori	Multiple Support apriori
MSE	Mean Squared Error
NMAE	Normalized Mean Absolute Error
NRMSE	Normalized Root Mean Squared Error
PFCM	Parallel Fuzzy C-Means
RBF	Radial Basis Function
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RMSE	Root Mean Squared Error
RS	Random Search
SD	Support Difference
SFS-SR	Sequential Feature Selection based on Stepwise Regression
SR	Stepwise Regression
SVM	Support Vector Machine
TID	Transaction ID
UCI	University of California, Irvine of machine learning
WFS	Weighting Feature Selection
XB	Xie and Beni

CHAPTER ONE

1 INTRODUCTION

1.1 Background to the Research Problem

In recent years, the rapid growth and large volume of data increasingly requires developing a prediction model based on data mining and knowledge discovery techniques. However, the prediction model is able to extract knowledge from a database and deal with different application domains for estimating a future value. Consequently, such a model can be used to improve prediction and decision making.

Knowledge Discovery in Databases (KDD) originating from the AI research field, is the sequential process (phases) of identifying a hidden pattern in a large database, which can be accomplished by developing a model or by integrating some techniques for useful information (knowledge) discovery (Fayyad et al., 1996a, Fayyad, 1996). KDD performs several phases or steps for extracting high-level knowledge from low-level data that is significant in decision making and human support (Mitra et al., 2002).

KDD is a rapid growing field in which Data Mining (DM) plays a pertinent role by using a number of techniques including a statistical approach, Artificial Neural Network (ANN), Fuzzy Logic (FL), rule induction, evolutionary algorithm and graphic visualization (Shaw et al., 2001, Abonyi et al., 2005).

Indeed the association rules mining is one of the most important tasks used in DM, which can be applied in different domains. Association rule discovery has been widely studied throughout the state-of-the-art techniques (Liu et al., 1999, Kiran and Reddy, 2010a). Association rules discovery was presented in Agrawal et al. (1993), which intends to extract the characteristics, hidden association pattern and the correlation between the items (attributes) in a large database (Kannan and Bhaskaran, 2009, Kiran and Reddy, 2010a).

Apriori algorithm is a classic and most popular algorithm for association rules extraction from a database, which was developed by Agrawal and Srikant in 1994 to extract strong rules (knowledge) of highly frequent itemsets in a transaction database using the pre-defined threshold measures, minimum support (*minsupp*) and minimum confidence (*minconf*). Association rules are formally written and presented in the form of "IF-Then" as follows: $X \rightarrow Y$, where X is called the antecedent and Y is called the consequent. For instance, "if the customer buys milk then he/she buys cornflakes", the antecedent is "buys milk" and the consequent is "buys cornflakes". The *minsupp* is the occurrence (frequency) of X and Y together, $\mathrm{support}(X \cup Y)$, and *minconf* is the ratio of occurrence (X and Y) divided by (/) occurrence (X) (i.e. $\frac{\mathrm{support}(X \cup Y)}{\mathrm{support}(X)}$).

One of the advantages of the association rule discovery is extracting explicit rules that are practically important for the user/human expert to understand the application domain. Thus, this can be facilitated to adjust (extend) the rules manually with further domain knowledge, which is difficult to achieve with other mining approaches (Gedikli and Jannach, 2010).

In this thesis a knowledge discovery model for prediction by employing fuzzy clustering techniques, association rules mining approaches and feature selection method is developed. This model provides a base for the KDD and DM technologies to extract knowledge from different data sets in realistic domains. The effectiveness of the proposed prediction model is evaluated using different data sets of different application domains as follows: (i) two case studies in road traffic domain, (ii) different benchmark data sets from UCI and KEEL repositories and (iii) benchmark data set of the gas furnace time series.

1.2 Research Motivation and Justification

Srikant and Agrawal (1996) introduced the problem of extracting association rules from quantitative attributes (numeric data set) by using the partitions method for these attributes (Srikant and Agrawal, 1996). Some of the current association rule mining approaches for quantitative data neglect the values of the intervals boundaries of the partitions. This causes sharpness for the boundary intervals which cannot reflect the intuitiveness of human understanding, justifiably argued by (Kaya and Alhajj, 2003, Kuok et al., 1998). Instead of using partitions method for the attributes, it is better to adopt the advantage of fuzzy set theory in a smooth transition between fuzzy sets. As a whole, the fuzzy approach is used for transforming quantitative data into fuzzy data. Fuzzy Logic (FL) is capable of handling the uncertainty by using fuzzy sets to perform the human way of thinking and reasoning (Mitra et al., 2002).

A variety of approaches have been developed in order to extract fuzzy association rules from quantitative data set (Hong et al., 2004, Zhang et al., 2005, Huang et al., 2006). These rules are presented in the form of "IF-Then" statements. These approaches assume that the member functions of fuzzy set are known in advance, or use an expert to define fuzzy sets of a quantitative data. The problem of association rules extraction from a quantitative data is investigated in this thesis using fuzzy clustering techniques. Fuzzy clustering is a suitable method to transform quantitative data into fuzzy ones, taking the advantage of fuzzy set theory over the partition method concerning the smooth transition among fuzzy sets. Fuzzy Association Rules (FARs) mining is adapted in this research as a solution for extracting knowledge from the quantitative database.

As mentioned earlier, the association rule mining aims at discovering the relationship (rules) among the data attributes (features), which depends on *minsupp* and *minconf* (Liu et al., 1999, Kiran and Reddy, 2009). Consequently, large numbers of rules are anticipated, particularly if *minsupp* is set to be very low. Practically, single *minsupp* is a vital parameter that controls the extracted number of association rules (Hu and Chen, 2006). Conventional association rule mining approaches like Apriori (Agrawal and Srikant, 1994) and Frequent Pattern-Growth (FP-Growth) (Han et al., 2000) are based on a single minimum support threshold. However, it was observed that using single *minsupp* causes a dilemma called “rare item problem” as follows (Liu et al., 1999, Hu and Chen, 2006, Kiran and Reddy, 2009):

- Frequent itemsets that contain rare items are missed when *minsupp* is set too high.
- Large numbers of frequent itemsets are produced when *minsupp* is set to a very low value, i.e. a very low *minsupp* value leads to the generation of all possible combinations (combinatorial explosion).

To solve the rare item problem Liu et al. (Liu et al., 1999) first developed the Multiple Support model called Multiple Support Apriori (MSapriori) algorithm. MSapriori is based on the idea of setting a Minimum Item Support (MIS) for each item in a database, i.e. employing multiple *minsupp* for different items in database, instead of using single *minsupp* for whole database. Hence, MSaproiri is expressed as a generalization of Apriori algorithm. Almost different MIS values can be assigned to assess different frequent items. As a result, this model facilitates the generation of frequent itemsets of rare items and prevents the production of uninteresting frequent itemsets (Hu and Chen, 2006). More recently, an approach has been developed to improve MSapriori called Improved Multiple Support aproiri (IMSapriori) (Kiran and Reddy, 2009).

In other words, using single *minsupp* approach considers the items (attributes) of a data set having the same frequencies. However, the nature of real-life applications is that it includes data items of different frequencies. In a supermarket transaction database that contains many items, most of these items are frequent while others are rare. Typically, there are many products (goods) of low price while others are high. Thus, buying the frequent products of low price seems to be higher than the infrequent (rare) products of relative high price, which reflects on a transaction database (i.e. a transaction database consists of both frequent and rare items). Consequently, useful knowledge can also be offered from the frequent itemsets that contain rare items (Hu and Chen, 2006, Kiran and Reddy, 2010b). For example, the supermarket database contains frequent items such as milk and cornflakes, and it also includes rare items such as a mattress and pillow.

Multiple support approaches are adapted in this thesis in order to deal with the limitations of using single *minsupp* to extract important and significant knowledge (rules). Furthermore, representative rules (diverse rules) are also extracted. In this thesis a post-processing method is proposed to select both of best and diverse rules. The diverse rules cover some training data of low frequency data attributes (i.e. clustering rules to classify and find a small set of significant rules in each cluster to be representative), which can be used later to cover particular cases.

Generally, association rule mining is an unsupervised technique, i.e. the consequent part of an association rule does not necessarily include the target attribute (non-predefined target). Associative classification (AC) is a distinctive case (special case) of association rules mining called Classification Association Rules (CARs), which integrates association rule mining and classification. Basically, AC is employing association rule mining for building a classification model. Since, CARs contain only the rules that hold a target attribute (class lable) in the right-hand side of a rule (consequent part of a rule), AC is formally expressed as $X \rightarrow C$, where X is called antecedent and C is called consequent and must be restricted to a labelled class (target attribute) (Thabtah, 2007).

In the past few years, AC approaches have received much attention with a significant amount of research being conducted successfully in the area of building accurate models for classification. Moreover, AC has been proved and confirmed to be a promising approach by achieving better results than conventional classification techniques (Kianmehr and Alhajj, 2008). Also, the output of AC is represented in the form of “IF-Then” rules; this is considered one of the main characteristics of using AC over conventional techniques. Therefore, the rules in AC are easy to understand and interpret by the user/human expert.

Typically, the association rule mining and AC approaches produce a large number of rules, particularly when *minsupp* is set to a very low value. Again *minsupp* is a key element to control a number of generated rules. When the *minsupp* is high, then usually it can be reduced from a number of generated rules. However, many of the important rules of high confidence values will be missed.(Thabtah, 2007).

Although, the exiting approaches and the reported studies have highlighted the effectiveness of using AC, there are several limitations and issues in building a classification and prediction model. Firstly, the number of generated rules is high. Secondly, the process of extracting the significant rules (useful knowledge) offers a challenge for building an effective and accurate model (Kianmehr and Alhajj, 2008). This thesis investigates the limitations of the current AC approaches, in particular, using single *minsupp*, using an objective measure (*minconf*), applying level-wise like Apriori fashion and generating non-dominating rules (see Section 4.4.1). These limitations are addressed through the proposed Fuzzy Associative Classification Rules (FACR) algorithm by adapting and integrating a very recent approach, specifically the improved multiple support, AC and vertical data format scan approaches. In addition, a post-processing method is employed in order to select significant and diverse rules.

One of the most important issues affecting on a prediction model error and performance is using a high dimensional data (all data attributes (features)) (see Chapter 5). Feature selection method provides the capability to handle this issue. In this thesis, a feature selection method namely Weighting Feature Selection (WFS) is developed to enhance the proposed prediction model. WFS is based on two mechanisms for selecting a subset of feature that represents a whole data set. WFS is able to improve the prediction model through minimizing a prediction error and reducing the number of generated rules.

The current models for prediction and decision making are fitting for a specific domain. Hence, the challenge is offering to design a comprehensive model in order to extract knowledge (pattern) and deal with various data from different application domains. The development of a generalized and an effective model for prediction is required through discovering significant patterns (knowledge) from a large database. Therefore, extracting a new pattern from a database for a future value prediction is an important goal in DM. It is considered that there is no optimal and best model for all data forms and application domains. Thus it is necessary for a reliable prediction model to be applicable in different application domains.

1.3 Research Aims and Objectives

The overall aim of this research is to develop effective and enhanced approaches for building a knowledge discovery model (prediction model) for predicting a future value accurately, which can be applied in different application domains. The outcomes of this research can be achieved through the design of comprehensive DM approaches (such as association rule mining) and Artificial Intelligence (AI) techniques (such as

Fuzzy Logic (FL)). Several novel approaches are designed, integrated, implemented and validated to extract reliable knowledge from databases needed to build a prediction model.

In summary, the main objectives of this research are:

- To investigate and review the literature in the area of DM and AI techniques to explore the state-of-the-art techniques (Fuzzy Logic, association rule mining, associative classification, feature selection method) and application domains, as well as to identify their strengths and weaknesses and highlight the current research challenges in knowledge base building.
- To identify existing DM and knowledge discovery techniques that can be implemented to build a prediction model. This includes fuzzy clustering algorithm to transform the quantitative data into fuzzy data, and Apriori approach to extract association rules from fuzzy data. In addition, to demonstrate the merits of associative classification approaches to facilitate the process of building a prediction model.
- To develop and validate an effective prediction model using an integration of fuzzy clustering algorithm, multiple support and associative classification approaches.
- To illustrate the merits of feature selection methods for reducing high dimensional data and develop an effective feature method that can significantly be reflected on the prediction power.
- To illustrate the capability and effectiveness of the prediction model by applying it to different case studies and benchmark data sets of different application domains to ensure generality of the model.

1.4 Research Contributions

The main contributions of this investigation are summarized as follows:

- Extensive literature review has been conducted for existing association rule mining, fuzzy association rules, associative classification approaches and feature selection methods with their application domains. In addition, benchmark evaluation techniques for the performance measurements for error calculations of the prediction model are explored.
- A knowledge discovery model for prediction is developed. This model is based on the integration of Fuzzy C-Means (FCM) to construct the fuzzy set and Apriori to extract the Fuzzy Association Rules (FARs). These FARs are used to build the knowledge base. The application of the developed prediction model has been demonstrated in two cases of road traffic management domain.
- The prediction model has been extended to utilize a multiple support concept to tackle with unbalanced database (i.e. the nature of data of the real-life applications contain attributes (items) of different frequencies, where the data set includes items of high frequency while others are rare). Furthermore, an effective diversification method is proposed to cluster the FARs (to discover both strong and representative (diverse) FARs across the rule space). This is based on employing the distance (dissimilarity) and the sharing function technique of multi-objective optimization.
- A hybrid prediction model is proposed entitled Fuzzy Associative Classification Rule Mining (FACRM), based on further development and improvement. The main characteristics of this new hybrid model are:

first, pruning directly for the FARs and improving the model performance; second, avoiding of uninteresting/insignificant (non-dominating) rules, thus, useful knowledge (rules) are extracted (frequent pattern that contains rare items); third, using the proposed diversification method. The hybridization can be achieved by integrating the recent and effective algorithms/approaches through; using Gustafson-Kessel (G-K) fuzzy clustering algorithm, utilizing the improved multiple support algorithm, adapting the vertical data format (scanning format for the fuzzy data), and employing the associative classification approach, in order to extract the frequent pattern that contains the rare items (attributes) and to limit the combinatorial explosion (restrict from uninteresting frequent itemsets including frequent items).

- A feature selection method Weighting Feature Selection (WFS), is proposed, which is based on two feature selection mechanisms, weight and intersection operators. These mechanisms are based on the integration of the common and well-known feature selection techniques. This method improves the prediction error and performance.

1.5 Research Process

Research is defined as a systematic process of steps that investigate gather and analyze information in order to increase our understanding and solve existing problems and issues (Creswell, 2005). The research philosophy as presented in (Saunders et al., 2007) is classified into positivism and phenomenology (or interpretivism) paradigms.

Positivism paradigm is a philosophical system based on experience and empirical knowledge of natural phenomena, which it is established in physical and natural science (Remenyi et al., 1998, Kumar, 2010). In the positivism paradigm the researcher is independent of what is being researched, therefore valid knowledge is only the observable and measurable phenomena (the research is conducted in the form of objective interpretation) (Collis and Hussey, 2003). This paradigm uses a quantitative approach, and adopts a scientific tradition approach. Positivism intends to identify, measure and validate any phenomena in order to prove the observations of scientific justification and to provide a logical explanation. The quantitative research involves collecting numeric data, and then quantifies and measures this data. As a result, the data is analysed and presented based on mathematical and statistical perspectives (Charles, 1995, Kumar, 2010). Quantitative research examines natural phenomena developed in the natural sciences (Myers, 1997).

The main research methodologies (research strategies) used in a quantitative approach are summarized as follows: surveys, experimental studies, formal methods, longitudinal studies, and cross-sectional studies (Creswell, 2003, Myers, 1997).

The phenomenology paradigm considers phenomena as objects of observation. This paradigm depends on human behaviour, experience and understanding to explain and interpret the phenomena, i.e. the phenomena is explained based on human subjectivity (the research is conducted in the form of subjective interpretation) (Remenyi et al., 1998, Myers, 1997). This paradigm uses a qualitative approach to support an understanding of the social context (Creswell, 2003). A qualitative approach examines social and cultural phenomena developed in social sciences (Myers, 1997), focusing on non-numerical data (textual). This approach was characterized by using an open-ended format that makes flexibility in analysis (Lancaster, 2005).

The main research strategies used in a qualitative approach are summarized as follows: case studies, action research, ethnography (participant observation), and grounded theory (Creswell, 2003, Myers, 1997).

The research approaches (reasoning process or methods of reasoning) are divided into two main approaches, namely deductive and inductive (Trochim, [accessed September 2011]). The deductive approach works in a “top-down” fashion through four steps. Firstly, a theory concerning a particular research is developed. Secondly, a hypothesis or hypotheses are formed. Thirdly, the hypotheses are addressed by gathering of observation. Finally, the hypotheses with the use of particular data are validated. The inductive approach follows a “bottom-up” method consisting of four steps. Firstly, the observations are made. Secondly, patterns and regularities are identified. Thirdly, tentative hypothesis or hypotheses are formulated. Finally, a theory or general conclusion is developed and drawn based on the analysis.

Generally, the research conducted in this thesis based on positivism paradigm, which acts the following procedures (processes) for each chapter: (i) identifying the problems and issues based on gathered information and background research, (ii) designing and proposing a solution(s) to tackle the defined problems or issues , (iii) evaluating, validating and testing the proposed solution empirically using a number of benchmark data sets in different applications domains, (iv) analyzing and comparing the results with other well known techniques/ models in order to verify the effectiveness of the proposed solution/ model.

The systematic process of steps applied for the research presented in this thesis are as follows:

- The issues, problems and challenges of the current techniques are identified, and then an initial (prototype) prediction model based on the well-known approaches is proposed in the first stage.
- Further improvement is carried out in order to overcome the limitations of the first stage. The outcomes of this second stage are the foundations of a proposed prediction model.

- An incremental enhancement in order to improve the model developed is the third stage.
- A feature selection method is proposed to enhance the performance of the proposed prediction model through reducing the prediction error value and minimizing the number of generated rules.
- Different data sets are used to validate the proposed prediction models in Chapter 3 and Chapter 4. In addition, the data sets used in Chapter 4 are also applied to test the proposed feature selection method in Chapter 5. It is worth mentioning that the source of all these data sets and methodology are well referenced. These data sets are publically available and anonymous (not belonging to private or specific individual(s)), which are collected from simulation models (such as road traffic data) and other benchmarks data sets.
- The performance of the proposed model has been compared with well-known and exiting techniques, in order to refine the capability and effectiveness of the proposed model.
- The analysis is performed, which finds that the proposed model is comparable with other well-known techniques.

1.6 Statement of Ethics

The Harvard referencing (Harvard Bradford style) method has been used to give credit to individual(s) and organization(s) whose work has been used in this study. The source of all data and information (used directly or indirectly) is clearly declared in this study. This research has used existing data sets (benchmark data sets) already in the public domain where issues of anonymity and confidentiality have already been dealt with. Also, there are no issues regarding purposeful collection of data and obtaining informed consent. Furthermore, the research does not involve any commercially sensitive collaborations.

1.7 Thesis Organization

The rest of the thesis and remaining chapters are organized as follow:

Chapter 2 outlines a review of the literature related to association rules, fuzzy logic, fuzzy clustering techniques, fuzzy association rules and similarity. Associative classification approaches, vertical data representation, Artificial Neural Network (ANN), Support Vector Machine (SVM), Stepwise Regression (SR), Classification and Regression Tree (CART) and feature selection method are critically evaluated and the common evaluation criteria are analysed.

Chapter 3 implements and evaluates two knowledge discovery models for prediction. The first model, integrates FCM and Apriori approach for the purpose of performing prediction in two cases in a road traffic domain and the limitations of this model are critically reviewed. The second model proposes an approach called Diverse Fuzzy Rule Base (DFRB). DFRB is able to extract knowledge based on multiple support approach. DFRB selects the best and diverse rules for building a reliable prediction model, which is applied in a road traffic domain and other benchmark data.

Chapter 4 presents the proposed hybrid model for prediction. The proposed model is evaluated using two sets of experiments validation illustrated in the experimental results section thus demonstrating how previous work on prediction is enhanced and improved by the proposed model.

Chapter 5 employs two mechanisms for feature selection for developing a feature selection method. The proposed Weighting Feature Selection (WFS) method is validated through a number of experiments and comparative analysis.

Chapter 6 summarizes the general conclusions and main contributions of the research conducted in this thesis. Recommendations for future work and further research directions are proposed.

1.8 Publications

The following publications have been resulted in from the research conducted for this thesis:

- Sowan B., Dahal K.P. and Hossain A.M., "Fuzzy Association Rule Mining Approaches for Enhancing Prediction Performance-A Comparative Study", A journal paper submitted to the International Journal on Artificial Intelligence Tools (IJAIT), Under review, 2011.
- Sowan B.I., Dahal K.P., Hossain A.M. and Alam M.S., "Diversification of Fuzzy Association Rules to Improve Prediction Accuracy", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010) IEEE World Congress on Computational Intelligence (IEEE WCCI 2010), Barcelona, Spain, pp 1202 -1209, 2010.
- Sowan B., Dahal K.P. and Hossain M.A., "Fuzzy Multiple Support Associative Classification Approach for Prediction ", Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2010), L. Rutkowski et al. (Eds.): Part I, Springer Verlag, Lecture Notes in Artificial Intelligence (LNAI), vol. 6113, Poland, pp. 216–223, 2010.

- Sowan B., Dahal K.P. and Hossain M.A., "Knowledge Discovery based on Integrated Fuzzy and Apriori Approach for Prediction", Proceedings of 3rd International conference on Software, Knowledge, Information Management and Applications (SKIMA 2009), Fez, Morocco, pp. 70-77, 2009.

CHAPTER TWO

2 REVIEW OF EXISTING TECHNIQUES AND EVALUATION CRITERIA

2.1 Chapter Overview

The main purpose of this chapter is to present a literature review which focuses on several relevant themes: (i) a general review of Knowledge Discovery in Databases (KDD) and Data Mining (DM), (ii) a more extensive review of the literature related to various association rule mining approaches such as, Apriori, AprioriTid, Frequent Pattern Growth (FP-Growth) and Multiple Support algorithms, (iii) a brief background on Fuzzy Logic (FL), clustering techniques in particular fuzzy clustering algorithms and Fuzzy Inference System (FIS) (iv) an overview of the Fuzzy Association Rules (FARs) approach and the work which has been conducted in the literature, and (v) similarity of association rules, associative classification approaches, vertical data representation, Artificial Neural Network (ANN), Support Vector Machine (SVM), Stepwise Regression (SR), Classification and Regression Tree (CART) and feature selection method. Finally, the common and benchmark evaluation criteria are discussed.

2.2 Knowledge Discovery and Data Mining

The relative and iterative Knowledge Discovery in Databases (KDD) steps that were described in (Fayyad, 1996, Fayyad et al., 1996b, Mitra et al., 2002) are shown in Figure 2.1 and include:

- An understanding of a problem and its application domain increasingly requires a domain background (a prior knowledge) and to determine the purpose of this application.
- A collection of a target data set (sample of data, focusing on a subset of variables).
- A preparation, pre-processing and reduction of data (cleaning data, which ensures a completion of a data records, and removing noisy data). The reduction of data is achieved by selecting a feature subset that represents the data.
- An application of Data Mining (DM) techniques (for instance, association rules, classification, regression, clustering, etc).
- An evaluation of a discovered knowledge (pattern interpretation and understanding the result).
- A use of extracted knowledge in a decision making process.

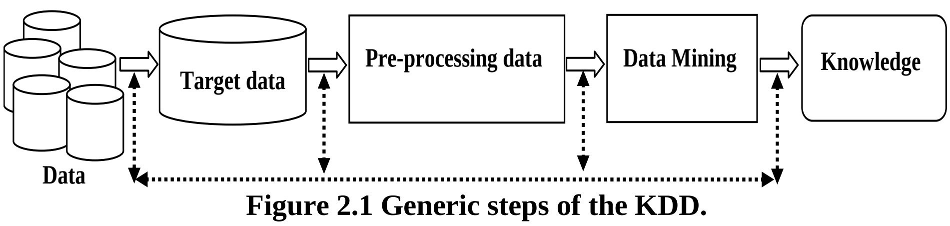


Figure 2.1 Generic steps of the KDD.

The iterative dependency shown in Figure 2.1 indicates how any change in any step will have an effect on the other processes.

2.2.1 Data Mining Tasks

DM is one of the essential and vital phases for discovering knowledge in KDD. DM is concerned with building models capable of finding useful surprising patterns, unknown, non-trivial relationships between attributes (extract knowledge) and interesting rules in relatively large databases. (Kanellopoulos et al., 2007).

DM, also known as data analysis technology (Hand et al., 2001, Fayyad et al., 1996a), focuses on algorithms that discover useful knowledge from volumes of data after pre-processing.

KDD and DM are considered as knowledge management technologies for knowledge production, which deal with that knowledge (rules) once extracted, in order to manage, integrate and verify it. This knowledge helps in the decision making process and improves competitive advantage for organisations (Liyanage et al., 2009). The importance of knowledge that already exists in the Knowledge Base (KB) reflects the quality to predict the future value. Prediction is a knowledge intensive process that requires superior knowledge management to produce an appropriate process effect. As a result, knowledge management is applied to verify knowledge, thus its predictions provide support for the application domain (Shaw et al., 2001).

Useful knowledge is an ultimate goal and can be determined through an interactive process of KDD (Fayyad et al., 1996b). Furthermore, DM aims to discover useful, comprehensive and interesting (interpretable) knowledge to act as the human expert.

DM is an important step of the KDD process that indicates the strength of KDD as shown in Figure 2.1. DM is widely used in different application domains such as: e-commerce, e-learning, road traffic, medicine, marketing and other domains (Imberman, 2002, Menon et al., 2005, Romero and Ventura, 2007, Marukatat, 2006, Lei and Ren-hou, 2007).

DM is involved in predictive and descriptive models that are applied in many different tasks as shown in Figure 2.2 (Dunham, 2002).

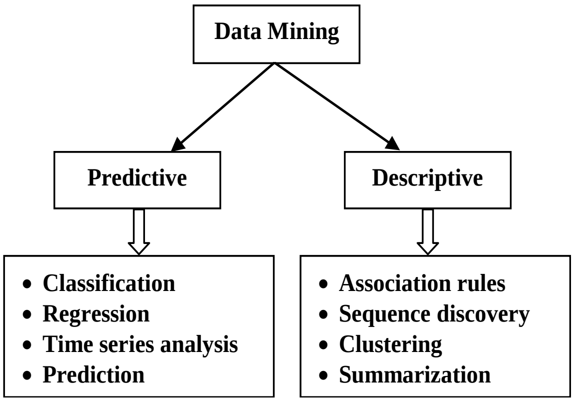


Figure 2.2 Data mining tasks.

The predictive model is concerned with an identification of the results obtained from various volumes of data for predicting (forecasting) the future values of a new data. For example, predicting the future sales for a customer depends on his/her historical data such as: age, gender and purchase items. The predictive model includes the following tasks (Bose and Mahapatra, 2001, Dunham, 2002, Romero and Ventura, 2007):

- *Classification*: splitting a data set into predefined groups or classes. For example, in airport security, criminals can be recognized by scanning each passenger's face in a way that the facial expressions are divided into patterns (face shape, distance between eyes, blinking eyes, eyebrows, etc), these patterns are then compared with the existing training sets in the database. Classification aims at building a model from different data attributes, and one of these attributes is the target attribute (class). Such a classification model is able to predict the class of a new case (set of attributes).
- *Regression:* fits a data set to an equation and works well with continuous quantitative data. For example, it uses the linear equation ($y = ax + b$) and determines a suitable value for a and b with a given value of x to

predict the future value of y (Chapple, [accessd February 2008]), i.e. suppose that the class (dependent variable) y is a combination of different attributes x_i (independent variables) associated with a coefficient a which can be calculated from the training data. In other words, the regression is a special case of the classification task when the target attribute is numeric.

- *Time series analysis*: the sequence of observations of well-defined data items (attributes). This sequence is examined through time (hourly, daily, weekly…etc) to predict the future value of the time variable. For example, the purchase stock can be predicted from *x* company based on testing a period of one-month.
- *Prediction*: tends to find out the future value based on the current or previous or old (historical) data. For example, river flooding prediction can be achieved based on various factors, such as: water level, rain amount, humidity, etc.

The descriptive model can recognize the relations in data with the aim of studying and focusing on the characteristics of data. The descriptive model includes the following tasks (Bose and Mahapatra, 2001, Dunham, 2002, Romero and Ventura, 2007):

- *Association rules*: the practice of discovering the association and the correlation between attributes (items) in a large database. Association rules mining is able to find the frequent itemsets (itemsets include items that have high frequency in a database) and the well-know application of the association rules which is the market basket analysis. For instance, a customer who buys a keyboard may buy a mouse at the same time.
- *Sequence discovery*: tends to find a sequence event (pattern) in a data set. For example, if a customer buys a camera, within three months they will

buy photographic supplies and within six months an accessory item (60% of customers who first buy X also purchases Y within two weeks).

- *Clustering*: splits the data sample (observations) into meaningful groups (non-predefined groups) according to some criteria. The difference between classification and clustering is that the former is supervised learning consisting of a set of labelled patterns which are able to classify (label) a new unlabelled data. Whereas the latter is grouping a set of data observations (unlabelled) into clusters where the labels are derived and linked with clusters (data driven) (Jain et al., 1999).
- *Summarization*: provides significant methods that derive useful information from a database to be representative (called characterization or generalization). For example, the value of the mean or standard deviation to represent the attributes of the data.

2.3 Association Rule Mining

Association rule mining is one of the most important tasks in DM, which plays a vital role in business applications such as, market basket for producing strong rules from frequent items in a transaction database (Hahsler et al., 2005). The problem of mining binary (Boolean) association rules from a database was introduced by Agrawal et al. (1993). Binary association rules mining is represented by either 1 or 0, if an item exists in a database, then the value is 1. Otherwise, the value is 0. The following is the basic definition for the association rule mining applied in a database.

Definition: Let $\mathrm{I} = \{\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_\mathrm{n}\}$ be a set of distinct items (attributes), and any set of items is called an itemset. Let $\mathrm{D} = \{\mathrm{t}_1, \mathrm{t}_2, \dots, \mathrm{t}_\mathrm{m}\}$ be a set of transactions database TID. Each transaction TID in D is formed from a set of items in I. A strong association rule is defined in accordance to the form of $(X \rightarrow Y)$ where, $X, Y \subset I$ and $X \cap Y = \emptyset$ approve *minsupp* and *minconf*.

Association rules are constructed as an antecedent (left-hand side) X and a consequent (right-hand side) Y. The association rules are extracted from a database based on two metrics (interesting measures): (i) Minimum Support Threshold value (*minsupp*) the time frequency for particular item(s) divided by the number of transactions, as given by Equation 2-1 and (ii) Minimum Confidence Threshold value (*minconf*) and *minconf* the percentage value for the support value of the antecedent and consequent together divided by the support value of the antecedent part, as given by Equation 2-2. An example of the calculation of support and confidence values and their equations for a rule "a customer who buys chips and milk, may also buy apples at the same time" is given below. The support value is equal to 33% and the confidence value for the previous rule is 66%. Table 2-1, Table 2-2 and Table 2-3 illustrate the Boolean association rules.

$$\text{Support} = \frac{X \cup Y}{\text{No} - \text{of} - \text{TID}} \tag{2-1}$$

$$\text{Confidence} = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \tag{2-2}$$

$$\text{Support} = \frac{\text{Chips} \cup \text{Milk} \cup \text{Apple}}{6} = \frac{2}{6} = 0.33$$

$$\text{Confidence} = \frac{\text{Support}(\text{Chips} \cup \text{Milk} \cup \text{Apple})}{\text{Support}(\text{Chips} \cup \text{Milk})} = \frac{0.33}{0.5} = 0.66$$

Table 2-1 Transaction items.

TID	Items
1	Chips, Milk, Apple
2	Chocolate, Juice
3	Chicken, Potato, Tomato
4	Chips, Juice, Milk, Apple
5	Chips, Milk
6	Chips, Lemon, Banana

Table 2-2 Boolean transaction items.

TID	Chips	Milk	Apple	Chocolate	Juice	Chicken	Potato	Tomato	Lemon	Banana
1	1	1	1	0	0	0	0	0	0	0
2	0	0	0	1	1	0	0	0	0	0
3	0	0	0	0	0	1	1	1	0	0
4	1	1	1	0	1	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	1	1

Table 2-3 Association rules.

Association rule	Support	Confidence
Chips, Milk → Apple	33%	66 %

Association rule mining can be divided into two main parts (Agrawal et al., 1993):

- Frequent itemsets are found if the support value of the itemset is greater than or equal to *minsupp*.
- Association rules are extracted if the confidence value of the frequent itemset is greater than or equal to *minconf*.

Apriori algorithm is one of the most well-know association rule mining algorithm. It is used to extract the rules and implicit knowledge from database attributes (field) in order to find the association and correlation between these attributes (Ye and Chiang, 2006).

Indeed, a database includes not only binary attributes, but also contains quantitative attributes. In the case of a quantitative data set the classical algorithms are not able to extract association rules from it directly. Hence, a fuzzy approach aims at converting the quantitative data set into fuzzy sets for its simplicity and flexibility to soften the interval partition boundaries of the attributes (Lu et al., 2003b). This flexibility of fuzzy sets is common seen in data mining in different areas such as: staff performance analysis (Huang et al., 2006).

Several algorithms have been applied to extract association rules such as: Apriori, AprioriTid (Agrawal and Srikant, 1994), Frequent Pattern Growth, in short FP-Growth (Han et al., 2000) and Multiple Support approaches and will be discussed in the next sub-sections.

2.3.1 Apriori Algorithm

Apriori algorithm is the simplest and most popular algorithm for discovering association rules from a large database (Agrawal and Srikant, 1994). Apriori algorithm works using two main steps as described in Figure 2.3 :

- Frequent itemsets are called (L_k), these itemsets are greater than or equal to *minsupp*.
- Candidate itemsets are called (C_k), and generated from (L_{k-1}). It is required to scan the original database for each itemset in the candidate generation step, in order to calculate the support value (scan the original database in C_1, C_2, and C_3 as it is described in Figure 2.4).

Apriori algorithm as shown in Figure 2.4 (steps 1 to 8, assuming that minsupp =2) has been developed by Agrawal and Srikant (1994) to extract the association rules from a transaction database. For example, a customer who buys a keyboard and monitor also buys a mouse (60% of all customers who purchase X and Y also buy Z). A detailed explanation about association rules and Apriori algorithm can be found in (Agrawal et al., 1993) and (Agrawal and Srikant, 1994) respectively.

L_k: Large frequent itemset of size k.
C_k: Candidate itemset of size k (generated from L_{k-1} based on join L_{k-1} with L_{k-1}).

L_1= {large frequent of the one item};
For (k = 2; $L_{k-1} \neq \emptyset$; k++) **do**
    C_k = candidates generated from L_{k-1}(join L_{k-1} with L_{k-1});
      {
        Insert into C_k
        Select $p.item_1, p.item_2 \dots p.item_{k-1}, q.item_{k-1}$
        From L_{k-1} called (p), L_{k-1} called (q)
        Where $p.item_1 = q.item_1 \dots p.item_{k-2} = q.item_{k-2}, p.item_{k-1} \neq q.item_{k-1}$
      }
    Foreach itemset c ∈ C_k do
        Foreach $(k-1)$ subset s of c do
            IF ($s \notin L_{k-1}$) then
                Delete c from C_k
            EndIF
        EndFor
    EndFor
    For each transaction t in database **do**
        Increment the count of all candidates in C_k that are contained in t.
        L_k = contains the candidates in C_k that is greater than or equal *minsupp*.
    EndFor
EndFor
The answer = $\bigcup_k L_k$

Figure 2.3 Apriori algorithm (Agrawal and Srikant, 1994).

Apriori algorithm example of Figure 2.4 is explained as follows:

- Database in step 1 is scanned to calculate the support value for each item, and then the items are stored as a candidate itemsets C_1 with their support values as shown in step 2.
- Candidate itemsets C_1 are moved to the frequent itemsets L_1 in step 3 if their support values are greater than or equal to *minsupp*.
- Frequent itemsets L_1 are joined up with L_1 to generate candidate itemsets C_2 in step 4 and step 5. Each candidate itemset is checked based on every sub-itemset which should be frequent itemset in the previous frequent itemsets L_1. Therefore, the support value for each candidate itemset C_2 is calculated based on scanning the database in step 1.
- Candidate itemsets C_2 are moved to the frequent itemsets L_2 in step 6 if their support values are greater than or equal to *minsupp*.

- Frequent itemsets L_2 are joined up with L_2 to generate candidate itemsets C_3 in step 7 and step 8. Each candidate itemset is checked based on every sub-itemset and should be frequent itemset in the previous frequent itemsets L_2. Therefore, the support value for each candidate itemset C_3 is calculated based on scanning the database in step 1.

Step1

Database	
ID	Item
1	CDE
2	ACE
3	BDE
4	CD
5	BE

Step2

Candidate itemset C_1	
Itemset	Support
{A}	1
{B}	2
{C}	3
{D}	3
{E}	4

Step3

Large itemset L_1	
Itemset	Support
{B}	2
{C}	3
{D}	3
{E}	4

Step4

Itemset
{BC}
{BD}
{BE}
{CD}
{CE}
{DE}

Step5

Candidate itemset C_2	
Itemset	Support
{BC}	0
{BD}	1
{BE}	2
{CD}	2
{CE}	2
{DE}	2

Step6

Large itemset L_2	
Itemset	Support
{BE}	2
{CD}	2
{CE}	2
{DE}	2

Step7

Itemset
{CDE}

Step8

Candidate itemset C_3	
Itemset	Support
{CDE}	1

Figure 2.4 Apriori algorithm example.

2.3.2 AprioriTid Algorithm

AprioriTid algorithm was developed in (Agrawal and Srikant, 1994) and works as the Apriori approach with the modification in calculating the support value for each itemset towards candidate generation step. i.e. AprioriTid better than Apriori in saving the scanning time of the calculating support value for each itemset in the candidate generation step (Apriori is scanning for the database in each candidate generation step) (Hong et al., 2004). AprioriTid algorithm operates in accordance with the following steps:

- Candidate itemsets are called ($\bar{C}_k$), which is a copy of the whole database at the first step ($\bar{C}_k$=Database).
- Frequent itemsets are called (L_k), if these itemsets are greater than or equal *minsupp*.

- Candidate itemsets are called (C_k) which is generated from (L_{k-1}).
- Candidate itemsets are called ($\bar{C}_k$). This is generated from ($\bar{C}_{k-1}$) and (C_k). Actually, ($\bar{C}_k$) is produced from the candidate itemsets ($\bar{C}_{k-1}$) and these itemsets exist in the candidate itemsets (C_k). In addition, it stores these itemsets in ($\bar{C}_k$) based on transaction record (TID) as in the ($\bar{C}_{k-1}$).

The AprioriTid approach is not required to scan the original database for each candidate itemset generation step due to the use of ($\bar{C}$). To understand how AprioriTid works, Figure 2.5 presents an example for further demonstration of the AprioriTid algorithm (steps 1 to 10, assuming that *minsupp* =2).

As depicted in Figure 2.5, the support value for each candidate itemset in C_2 is calculated through the iterative process over $\bar{C}_1$ which also generates $\bar{C}_2$. Then, a pass is carried on over $\bar{C}_2$ to calculate the support values for the candidate itemsets in C_3, and C_3 produces $\bar{C}_3$.

Step1

Database	
TID	Item
1	ACDE
2	BCE
3	ABCE
4	BE

Step2

C_1	
Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	4

Step3

$\bar{C}_1$	
TID	Set of itemset
	{{A},{C},{D},{E}}
	{{B},{C},{E}}
	{{A},{B},{C},{E}}
	{{B},{E}}

Step4

Large itemset L_1	
Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	4

Step5

C_2	
Itemset	Support
{A,B}	1
{A,C}	2
{A,E}	2
{B,C}	2
{B,E}	3
{C,E}	3

Step6

$\bar{C}_2$	
TID	Set of itemset
1	{{A,C},{A,E},{C,E}}
2	{{B,C},{B,E},{C,E}}
3	{{A,B},{A,C},{A,E},{B,C},{B,E},{C,E}}
4	{{B,E}}

Step7

Large itemset L_2	
Itemset	Support
{A,C}	2
{A,E}	2
{B,C}	2
{B,E}	3
{C,E}	2

Step8

C_3	
Itemset	Support
{A,C,E}	2
{B,C,E}	2

Step9

$\bar{C}_3$	
TID	Set of itemset
1	{{A,C,E}}
2	{{B,C,E}}
3	{{A,C,E},{B,C,E}}

Step10

Large itemset L_3	
Itemset	Support
{A,C,E}	2
{B,C,E}	2

Figure 2.5 AprioriTid algorithm example.

2.3.3 Frequent Patterns Growth (FP-Growth)

Frequent Pattern Growth (FP-Growth), which was developed by Han et al. (2000), is one of the DM algorithms to extract association rules (frequent itemsets) from a database using the divide-conquer technique. FP-Growth candidate itemsets do not need to be generated compared with Apriori algorithm. A database in this algorithm is represented as a tree in compact shape. The FP-Growth approach explores the following steps (Han et al., 2000):

- Counting the database items and reordering the original database. This can be achieved through two actions. First, compute the support value for each item in the database, and keep the items that are frequent (if the support value for each item is greater than or equal to *minsupp*). Second, sort each single item in descending order based on the support value and in addition remove the item from the database in which the support value is less than *minsupp* (assuming *minsupp* = 2), as shown in step 1-3 of Figure 2.6.
- Building up Frequent Pattern Tree (FP-Tree) that starts from "null" and then scanning each sorted transaction database. This makes a path from the null node (tree staring point) to the last item in each sorted transaction database. FP-Tree contains item node, support value, and path link, as shown in step 4 of Figure 2.6.
- Discovering the frequent items from FP-Tree. This can be achieved due to the construction of a table containing the frequent items starting with the item node and its support value (from leaf to root), in order to find the conditional pattern base. Subsequently find all the combinations of

the frequent patterns to discover the association rules, as it is shown in step 5 of Figure 2.6.

Step1 Database.

ID	Items
1	ACDE
2	BCE
3	ABCE
4	BE
5	BC

⇨

Step2 Support value of items.

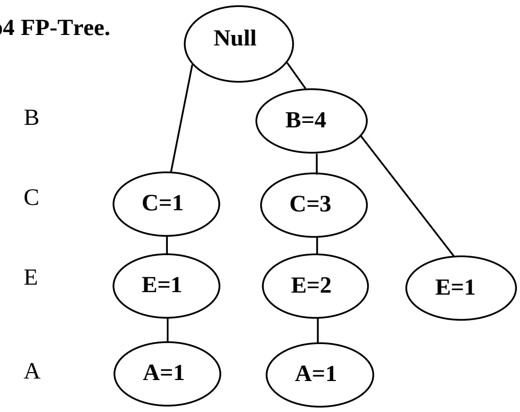
Item	Support value
B	4
C	4
E	4
A	2

⇨

Step3 Sorted database.

ID	Sorted items (descending order)
1	CEA
2	BCE
3	BCEA
4	BE
5	BC

Step4 FP-Tree.



Step5 Discovering FP.

Item (ascending order)	Conditional pattern base	Conditional FP-Tree	Frequent itemset
A	{(C:1 E:1), (B:1 C:1 E:1)}	{(C:2, E:2)}|A	(A,C),(A,E),(A,C,E)
E	{(C:1),(B:2 C:2),(B:1)}	{(B:3, C:3)}|E	(B,E),(C,E),(B,C,E)
C	{(B:3)}	{(B:3)}|C	(B,C)
B	ø		

Figure 2.6 FP-Growth example.

2.3.4 Multiple Support Approaches

Most of the association rules techniques have been extracted association rules based on the high frequency occurring. Conversely, natures of the real-life applications and their data sets are generally inconsistent which have both rare and frequent items. However, the rare items are difficult to be identified because of their low quantity data, which causes a rare item problem. To overcome the rare item problem, Liu et al. (Liu et al., 1999) proposed an algorithm called Multiple Support Apriroi (MSapriroi).

Example: In the supermarket transaction data set, the following items are infrequent such as: kettle and cooking pan, if the *minsupp* is set very low, then some other uninteresting itemsets will be extracted as follows:

Kettle → Cooking-Pan (minsupp=0.2%, minconf=0.6)
Bread → Pen (minsupp=0.2%, minconf=0.6)

The first rule is a rare itemset which is meaningful while its items are infrequent. The second rule is a non-sense, but because its items are frequent, that helps the items to be associated.

MSapriori is a generalization for Apriori algorithm in applying multiple minimum support thresholds, also called Minimum Item Support (MIS) (assign one minimum support threshold for each item) (Liu et al., 1999, Kanellopoulos et al., 2007). MIS is allocated for each item based on its frequency in the data and MIS can be defined in Equations 2-3 and 2-4 as follows:

$$\mathrm{MIS}(\mathrm{i}) = \begin{cases} \mathrm{m_i}\,, \text{if } \mathrm{m_i} \geq LS \\ \mathrm{LS}, \text{otherwise} \end{cases} \tag{2-3}$$

$$\mathrm{m_i} = \beta * \mathrm{f(i)} \tag{2-4}$$

where f(i) represents the actual frequency of each item i in the data, LS Least Support is user-defined, which is an assumed value as in traditional association rules, and β is a user-defined parameter to control the relation between MIS value for each item and its actual frequency which can be from 0 to 1.

In MSapriori each item is allocated an MIS. Thus, this can help in generating all frequent itemsets in case the support of an itemset satisfies its minimum MIS as described below.

Let $\mathrm{I} = \{\mathrm{i_1}, \mathrm{i_2}, \dots, \mathrm{i_n}\}$ be a set of items (attributes), and $\mathrm{D} = \{\mathrm{t_1}, \mathrm{t_2}, \dots, \mathrm{t_m}\}$ be a set of transactions database. Each transaction TID in D is formed by a set of items i in I where $\mathrm{i} \in \mathrm{I}$. The rule $\mathrm{R}, \mathrm{r_1}, \mathrm{r_2}, \dots, \mathrm{r_k} \rightarrow \mathrm{r_{k+1}}, \dots, \mathrm{r_n}$, where $\mathrm{r_k}, \mathrm{r_n} \in \mathrm{I}$, assuming that MIS is an

Minimum Item Support value for an item i, then the rule is accepted if its support value is greater than or equal to the minimum value of MIS from all items in each rule as shown below.

$$\min\left(\mathrm{MIS}(r_1), \mathrm{MIS}(r_2), \ldots, \mathrm{MIS}(r_n)\right) \leq \mathrm{support\_value}(R)$$

Traditional association rules algorithms, in particular Apriori algorithm, requires all subsets of a frequent itemset that should be frequent to have downward closure property. For example, let {milk, biscuit, coffee}is a frequent itemset in L_k; the subset of this itemset is as follows, {{milk, biscuit}, {milk, coffee}, {biscuit, coffee}}, all these subsets should be frequent itemsets in L_{k-1}. Conversely, MSapriori does not satisfy the downward closure property, yet it applies the sorted closure property.

It is observed that MIS has been calculated based on the β value. Therefore, it is necessary to set a proper value for β to generate frequent itemsets, including rare items, and to confine from producing uninteresting frequent itemsets including frequent items. In this manner, if β value is set to be high, then MSapriori suffers from generating the frequent itemsets including rare items. This is due to the approximate equality of MIS of the rare items and their support values as compared with frequent itemsets including frequent items.

Example: Consider the following items in a database, Bread, Pen, Kettle and Cooking-Pan. These items have support values 60%, 50%, 3% and 2% respectively. Let $\beta = 0.9$ and $LS = 1$, then MIS values for Bread, Pen, Kettle and Cooking-Pan are equal to 54%, 45%, 2.7% and 1.8% respectively. Therefore, the difference between support value of a rare (kettle) and its MIS is approximately very low (3% - 2.7% = 0.3%) as compared with frequent items (Bread) (60% - 54% = 6%). As a result frequent itemsets including rare items such as Kettle and Cooking-Pan will be missed. This is because the support values of the itemsets involving rare items are less than (not equivalent to) their MIS in particular, when the length of an itemset is increased, i.e. increasing the number of items to form one itemsets. For instance, Let the support value of an itemset including rare items {Kettle, Cooking-Pan} is equal to 1.5%. Therefore, this itemset will be missed due to its support (1.5%) being less than its MIS (1.8%) (min(MIS(Kettle),MIS(Cooking-Pan))).

In a different manner, if β value is set to be low, then it helps the support values of the rare items to be greater than their MIS. Also, it can be facilitated to set very low MIS values for the frequent items. Thus, the MSapriori leads to generating a large number of uninteresting frequent itemsets including frequent items. This is due to association of the items in all possible combinations.

Example: Continuing from the previous example let $\beta = 0.2$, then MIS values are 12%, 10%, 0.6% and 0.4% respectively. Thus, the difference between support value of rare items and their MIS values is higher than the previous values such as Kettle (3% - 0.6% = 2.4%). Also, this can increase the difference between support values of frequent items and their MIS values such as Bread (60% - 12% = 48%). As a result, low β value leads to generating uninteresting frequent itemsets, i.e. frequent itemsets including frequent items. These frequent itemsets have low support values. For instance, {Bread, Pen} is uninteresting frequent itemset including frequent items. Assuming the support value for {Bread, Pen} is equal to 12%, setting a low MIS value for this itemset which is equal to 10% (min(MIS(Bread),MIS(Pen))) by considering a low β value. Consequently, the support of {Bread, Pen} (12%) is greater than its MIS (10%). In other words, setting a low β value leads to generating uninteresting frequent itemsets (including frequent items) having low support values.

To solve the problem β value, Kiran and Reddy (Kiran and Reddy, 2009) developed an approach to improve MSapriori called Improved Multiple Support aproiri (IMSapriori). This approach uses a formula called Support Difference (SD) in calculation of MIS, which is an acceptable deviation (constant value) between an item support and its equivalent MIS. SD can be effectively able to generate frequent itemsets including rare items and to limit the combinatorial explosion (restrict from uninteresting frequent itemsets including frequent items). MIS is defined in this approach in Equations 2-5 and 2-6 as follows:

$$\mathrm{MIS(i)} = \begin{cases} \mathrm{m_i}, \text{if } \mathrm{m_i} \geq \mathrm{LS} \\ \mathrm{LS}, \text{otherwise} \end{cases} \tag{2-5}$$

$$\mathrm{m_i} = \mathrm{f(i)} - \mathrm{SD}, \mathrm{SD} = \lambda(1 - \alpha) \tag{2-6}$$

where, f(i) represents the actual frequency for each item i, LS Least Support is user-defined, which is an assumed value as in a traditional association rule, λ stands for statistical measure of the data such as: mean, median, mode, and maximum support of the items supports, α $(0 \leq \alpha \leq 1)$ is a control parameter. SD Support Difference ranges from 0 to λ.

The extraction of frequent itemsets $\mathrm{FIS(i)}$ can be found as follows:

$$\mathrm{support\_value(FIS(i_1, i_2, \dots, i_m))} \geq \min\,[\mathrm{MIS(i_1), MIS(i_2), \dots, MIS(i_m)}]$$

Generally, IMSapriori still used MSapriori approach in extracting the frequent itemsets (large itemsets). The differences between Apriori, MSapriori and IMSapriori algorithms are summarized in Table 2-4.

Table 2-4 Difference between Aprirori, MSapriori and IMSapriori.

Apriori	MSapriori	IMSapriori
Deal with one minimum support threshold for whole data set.	Deal with multiple item support for the data set, i.e. different MIS values for different items (attributes).	Deal with multiple item support for the data set by using Support Difference (SD).
Suffer from the rare item problem.	Solve the rare item problem by selecting an appropriate control parameter.	Solve the rare item problem and limit from combinatorial explosion.
Use downward closure property for pruning.	Use sorted closure property for pruning.	Use sorted closure property for pruning.

There have been several studies adapted FP-Growth (Uday Kiran and Reddy, 2009, Kiran and Reddy, 2010a) for mining association rules. An extension approach entitled Conditional Frequent Pattern Growth (CFP-Growth) was presented in (Hu and Chen, 2006) to improve the performance of the multiple support approach. This approach still suffers from pruning some of the rare items. More recently, a proposed approach called Improved Conditional Pattern Growth (ICFP-Growth) (Kiran and Reddy, 2010b), was employed in the improved multiple support approach. The experimental results demonstrated that the approach was able to generate the frequent itemsets involving rare items (rare association rules).

2.4 Fuzzy Logic (FL)

The term “Fuzzy Logic” was introduced by *Lotfi Zadeh* (1965), it is defined as "*a set of mathematical principles for knowledge representation based on degrees of membership rather than on crisp membership of classical binary logic*" (Negnevitsky, 2005). Thus, it is considered as one of the most important technologies viewed as a multi-valued logic based on fuzzy set theory. Fuzzy set deals with the approximate rather than precise methods in reasoning implication. In fuzzy set theory each fuzzy set or term has values on the universe of discourse and each value has a membership value that can range between 0 and 1 (Zadeh, 1988).

Moreover, FL offers reasoning qualitative mode (human way of thinking) as a style which mimics human decision making with logical expressions in the form of “IF–Then”, providing an intuitive method to describe the system drawn from human expressions (Russel and Norvig, 2003). In addition, reasoning in FL can be easily designed, understood and effortlessly and succinctly modified (Stathacopoulou et al., 2005, Aziz and Parthiban, [accessed December 2010]).

Fuzzy set is a function (shape) whose elements (x_i) on the universe of discourse have degrees of membership values (u_i) with possibilities of distribution in the range [0, 1]. Thus, a fuzzy set determines that each element (x_i) has a membership function in the interval of [0, 1]. In this form the fuzzy set can be represented as $\frac{U_i}{X_i}$ and the set of values as $\{\frac{u_1}{x_1}, \frac{u_2}{x_2}, \dots, \frac{u_n}{x_n}\}$. This is illustrated in Figure 2.7, where age is a fuzzy variable which has two fuzzy sets (young and medium) with a set of values (real number) on the universe of discourse (for example an age range from 20 to 50). These values are represented on the degree of membership values, where the universe of discourse defines a set of values that range between minimum and maximum bounds for the values of the fuzzy sets (Aziz and Parthiban, [accessed December 2010]).

As an example of the fuzzy set, by considering the membership values over the universe of discourse which are equal to $\{\frac{0.0}{25}, \frac{0.71}{30}, \frac{1}{32}, \frac{0.38}{45}\}$, as it is shown in Figure 2.7, this is a triangular function of highly condensed representation that covers some values in the discourse universe.

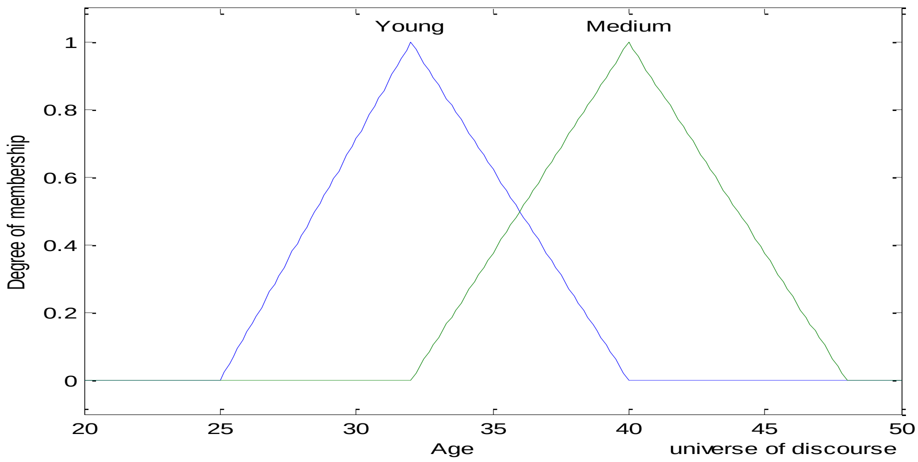


Figure 2.7 An example of fuzzy sets.

Fuzzy sets (terms) can be performed on the basis of a variety of operations such as: union, intersection and complement (Zadeh, 1965).

2.5 Clustering Techniques

The clustering is one of the data mining tasks. It aims to placing the data or individual objects into un-predefined groups, which conveys the notion of unsupervised learning (Xu and Wunsch, 2005, Matteucci, [accessed Novmber 2008]). In fact, the concept of the clustering data comes from the field of statistics that helps to carry the large data computation performance. Moreover, it is used in many fields, including document retrieval, pattern recognition, image segmentation and bioinformatics (Jain et al., 1999).

The classical meaning of clustering, assigns each data object to one cluster, whereas fuzzy clustering is unlike other clustering techniques and aims at assigning the data object to more than one cluster with different membership values (Shihab and Burger, 1998). That means that it is able to assign each data object into more than one cluster in a flexible manner, which is represented by the degree of membership value in each cluster. Most of the Fuzzy clustering techniques are based on an objective function, and the optimal clustering is determined by minimizing the objective function. The most popular fuzzy clustering techniques known as Fuzzy C-Means (FCM) and Gustafson-Kessel (G-K) algorithms are described below.

2.5.1 Fuzzy C-Means (FCM) Algorithm

Fuzzy C-Means (FCM) is one of the fuzzy clustering algorithms based on an objective functioning method, developed by Bezdek in 1981 adapting the fuzzy set theory, which assigns a data object (observation) to more than one cluster (Bezdek, 1981, Shihab and Burger, 1998).

FCM is commonly used for many applications because of its advantages in minimising objective function and finding the converge solution according to the fuzzy perspective which provides reliable membership values (Shihab and Burger, 1998). The FCM is considered better than any other clustering methods, such as k-means algorithm, because of its potential in distributing the data set observations (objects) into more than one cluster as a result of its flexibility, the membership value is categorized into more than one cluster (Kannan and Genova, 2005). Overall, a simple example on the FCM is shown in Table 2-5 and Figure 2.8 with two clusters.

Table 2-5 FCM example.

Two clusters

Data observation

X	Y	Cluster 1	Cluster 2
1.8	2	0.997	0.003
2	2.2	1	0
2	1.8	0.995	0.005
2.2	2	0.997	0.003
2	3.5	0.968	0.032
8.8	3	0	1
9	3.2	0.003	0.997
9	2.8	0.003	0.997
9.2	3	0.006	0.994
7	2.8	0.1	0.9

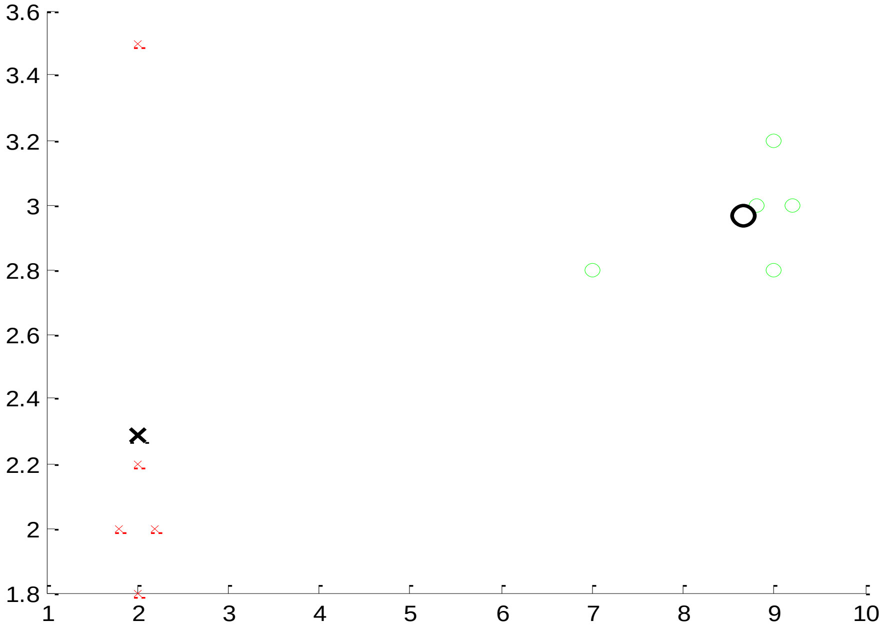


Figure 2.8 Two clusters of the data values from Table 2-5.

In the experiments in Chapter 3, FCM function is used as implemented in MATLAB 7.6, in order to determine the attributes centres (fields) of a crisp data set to find out the fuzzy data set. Four fuzzy sets are used in order to make a trade-off between the performance (efficiency) of FIS and prediction accuracy.

2.5.2 The Gustafson-Kessel (G-K) Algorithm

The Gustafson-Kessel (G-K) algorithm is a robust clustering method that can be applied in different domains including image processing and system identification. G-K algorithm realized Euclidean distance to Mahalanobis distance (Liu et al., 2008a) and enhanced the original FCM algorithm based on Euclidian distances. This enhancement has been attempted in order to identify different geometrical shapes of the clusters for one data set. In contrast with FCM, which is limited to spherical clusters shape (Balasko et al., 2008, Liu et al., 2009) and is affected by the outliers (Lesot and Kruse, 2006). The improved G-K clustering algorithm (Babuska et al., 2002) implemented in (Balasko et al., 2008) is applied in Chapters 4 and 5.

2.6 Fuzzy Inference and Expert System

Fuzzy Inference System (FIS) is a computer model that performs the reasoning based on fuzzy set theory and fuzzy rules of the “IF-Then” form. It is one of the most well-know applications of fuzzy logic used in different applications such as: decision analysis, expert system and prediction. Fuzzy inference can be defined as a method of mapping from a given input to an output through utilizing the fuzzy set theory (Negnevitsky, 2005, Guillaume, 2001).

The reasoning mechanism as shown in Figure 2.9 can be developed through the following steps: (i) entering the crisp input data (numeric data) into the system, where the input data is transformed into fuzzy data, (ii) adopting the knowledge rules already stored in the KB to perform the inference action by merging the fuzzy sets of all consequent parts (Then part in the “IF-Then” form) into one single fuzzy set (which stands for the output), and (iii) finally deriving and converting the output value into the crisp data. Thus, the rules are assessed.

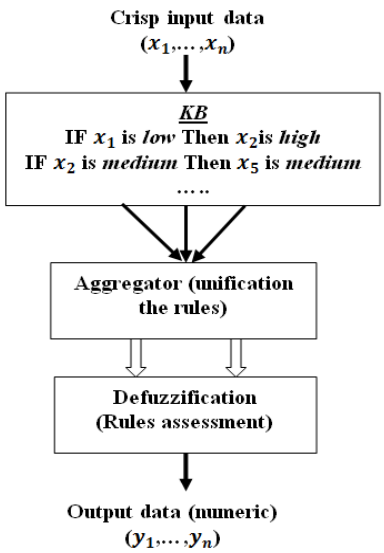


Figure 2.9 FIS diagram.

FIS is an effective model for prediction and developing the expert system through the following steps (Negnevitsky, 2005):

- Determining the problem domain (i.e. describing the data set) and identifying the fuzzy sets (linguistic terms) such as: low, medium and high.
- Creating the fuzzy sets by the use of a particular function such as triangular, trapezoid functions...etc.
- Extracting the fuzzy rules either by human experts or other techniques such as DM algorithms, after that storing these rules in the KB.
- Encoding the fuzzy sets and fuzzy rules to accomplish FIS, which carry out fuzzy inference into the prediction and expert system.
- Assessing and adjusting the system, i.e. to find out if the system fits the determined requirements at the beginning.

The experiments in this thesis used the FIS tool implemented in MATLAB 7.6, in order to help predict the future values.

2.7 Fuzzy Association Rules (FARs)

Fuzzy Association Rules (FARs) is a derived concept from the association rule when the association rule mining is employed, while the fuzzy approach is applied to deal with quantitative attributes (quantitative data or crisp one), representing them in a natural and understandable manner (Delgado et al., 2003, Chen et al., 2006a, Zhang and He, 2010).

A fuzzy approach is widely exploited among the intelligent systems, since it is very simple and similar to the human way of thinking (Mitra et al., 2002, Hong et al., 2004). It is used to transform quantitative data into fuzzy data through the identification of the membership functions. These membership functions of the fuzzy sets are defined in different methods, the basic one is based on human experience/know-how (Huang et al., 2006). However, in many instances, it is difficult to obtain the information required and to benefit from the human experience (Kaya and Alhajj, 2003). In the long run, a fuzzy clustering technique is commonly used for providing reliable membership values for constructing fuzzy sets.

Association rules and mining frequent patterns without a candidate generation step have been proposed in (Han et al., 2000), which used the data structure method called Frequent Pattern Tree (FP-Tree). This method works by first using data structure to condense the huge database into a smaller one. Second, it uses FP-Tree to build the database as a tree, in order to avoid the candidate generation step. Third, it uses the divide and conquer technique to partition the responsibilities of mining frequent patterns into a smaller one which it reduces from the search space of finding frequent itemsets.

The fuzzified quantitative attributes method have been considered in (Zhang, 1999) to overcome the partitions problem. The method is used the extended Equi-Depth Partition (EDP) (Srikant and Agrawal, 1996) known as Equi-Depth Partition Fuzzy

Terms algorithm (EDPFT). This algorithm was developed to extract association rules of the fuzzy terms, but using EDP in noisy (skewed) data sets is not useful.

An automated method for the identification of the fuzzy sets has been implemented in (Fu et al., 1998) to extract FARs. This method based on CLARANS (Clustering LARge Applications based on RANdomized Search) clustering algorithm, was applied in a real life quantitative data set. As a result, the use of the clustering methods leads to significant results instead of providing and identifying fuzzy sets based on user/human expert.

A Parallel Fuzzy C-Means (PFCM) has been proposed in (Xu et al., 2003) for mining FARs. The proposed algorithm works based on, the identification of the fuzzy sets while using PFCM algorithm and the extraction of FARs that satisfies *minsupp* and *minconf*. The parallel algorithm used the concept of a distribution system of 6 processors, copying the data set attributes in each memory that belonged to each processor. Their algorithm considered the number of data set records and concluded that using parallel algorithm is useful to scale up, size up and speed up the perspectives but that the increase of the data set attributes size will be unproductive.

A classification approach based on FARs called Classification Fuzzy Association Rule (CFAR) has been proposed in (Lu et al., 2003b) based on the integrated FCM and Apriori algorithm. It is applied on quantitative data known as Wine data set. Whereas the integrated approach has procured better results compared with other classification approaches such as C4.5 and Classification Based on Associations (CBA), this approach still applied single *minupp* threshold. It was found that, building a perfect classification model required a trade-off between producing fewer number of rules and achieving high accuracy.

The concept of mining multiple-level association rules (taxonomic concept) has been addressed in (Han and Fu, 1995) using top-down hierarchy rather than single level.

An example of the multiple-level (hierarchy tree) idea is illustrated in Figure 2.10, showing a tree, which is the root node at level 0, the inner nodes demonstrating categories (such as "Car") are at level 1, and the terminal nodes demonstrating models (such as "BMW") are at level 2.

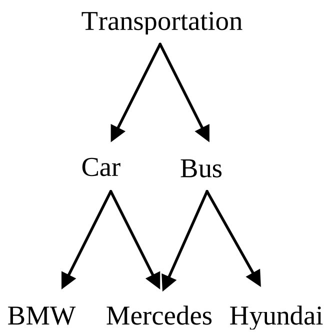


Figure 2.10 An example of multiple-level idea.

Also, the fuzzy generalized association rules technique (multiple-level) has been proposed in (Hong et al., 2003a). This technique used the fuzzy concept to convert the quantitative data set into fuzzy sets, considered as user predefined for the membership functions (assuming that the membership functions are known in advance). The paper suggests future work should use a proper technique for dynamically tuning the membership functions in order to avoid its acquisition bottleneck. Moreover, the problem of mining multiple-level fuzzy association rules (hierarchal concept) has been studied in (Hong et al., 2003b, Shitong et al., 2005). It was observed that, the previous techniques of multiple-level clearly used single *minsupp* threshold and assumed that the fuzzy sets were identified in advance. In addition, these techniques have been applied for a specific domain (taxonomy domain) such as form crops and supermarket data sets, otherwise, it cannot be implemented.

2.8 Similarity

A clustering technique for association rules was developed by Dechang and Xiaolin (Dechang and Xiaolin, 2008) based on the similarity of the antecedent part of the rules. The rules and its similarity coefficient are stored in the fuzzy simulation matrix $\mathrm{similarity}_{ij}$ to be understood by the user and then evaluated, as shown below. The author in (Dechang and Xiaolin, 2008) proved that Euclidian distance is not suitable to find the distance (dissimilarity) between the rules. The clustering technique proposed by (Dechang and Xiaolin, 2008) is worked as follows:

$$\mathrm{similarity}_{ij} = \begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nm} \end{bmatrix}$$

where $s_{ij} \in [0,1], 1 \le i \le n, 1 \le j \le m$, i, j represent the rules.

Assuming that five Fuzzy Rules (FR) are considered as follows:

FR1: IF x1= Low & x2=Low Then y=Low.
FR2: IF x1= Medium & x2=Low Then y=Low.
FR3: IF x1= Low & x2=Low & x3=Medium Then y=Low.
FR4: IF x1= High & x2=Medium & x3=High Then y=Medium
FR5: IF x1= High & x2=High & x3=Medium Then y=High

Fuzzy Coefficient Similarity (CS) is shown in Equation 2-7.

$$CS_{ij} = \frac{FR_i \cap FR_j}{FR_i \cup FR_j} \tag{2-7}$$

Then, CS values for all rules are as follow:

CS(FR1,FR2)=0.33, CS(FR1,FR3)=0.66, CS(FR1,FR4)=0, CS(FR1,FR5)=0, CS(FR2,FR3)=0.25, CS(FR2,FR4)=0, CS(FR2,FR5)=0, CS(FR3,FR4)=0, CS(FR3,FR5)=0, CS(FR4,FR5)=0.2.

However, in (Dechang and Xiaolin, 2008) the use of the Apriori algorithm for extracting these rules and the similarity of the antecedent parts only affects the result performance. Also, this technique does not take into account the distance between fuzzy sets included in FARs as depicted in Figure 2.11. The distance between Low and Medium is not equal to the distance between Low and Very High. For instance, let the distance between Low and Medium equal 0.33, then the distance between Low and High can be 0.66, and the distance between Low and Very High is equal to 1, in the distance range [0, 1]. This is reflected in Equations 4-1, 4-2 and 4-3.

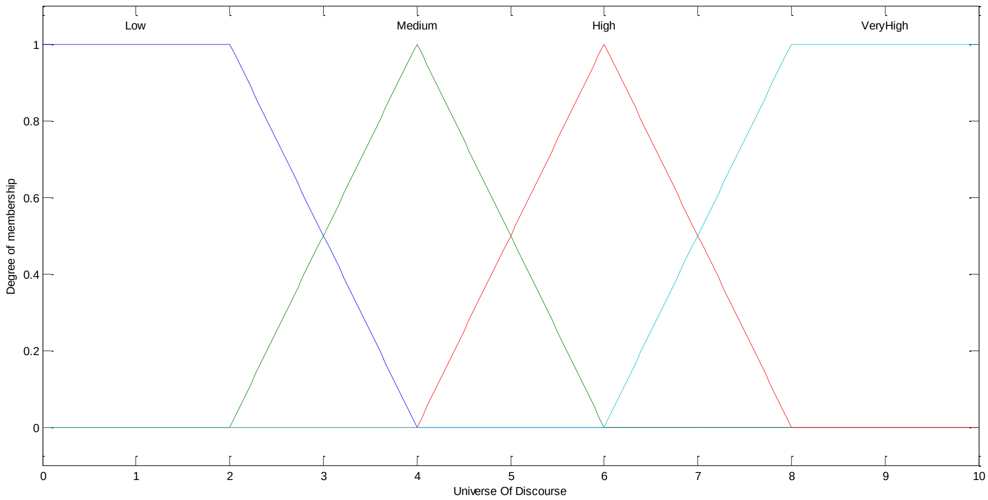


Figure 2.11 Fuzzy sets.

2.9 Associative Classification Approach

Associating rules mining is considered unsupervised learning since the rules contain all possible extracted rules and there is no specific output attribute (class label) after the “Then” part. A rule is accepted if it satisfies both of the *minsupp* and *minconf* thresholds. The classification rules (also called Associative Classification Rules or Classification Association Rules (CARs)) are regarded as supervised learning that be composed of the class label after the “Then” part of a rule. If the output attribute is discrete, then it is called classification. Otherwise, if the output attribute is continuous, then it is called regression. As a result, CARs is a special kind of association rules mining and it is used to build a predictive model. In other words, association rules mining aims at finding a correlation and relationship between the data attributes, while classification aims at allocating the data objects into a desired output value (Thabtah et al., 2006). Heuristic methods are adopted in most of the traditional classification approaches to find a small sub-set of rules. This results in missing many important rules that might be convenient in some other cases (Liu et al., 1998, Thabtah et al., 2006).

Classification is a well-known data mining task, many studies of different real-life applications have used the popular and widespread techniques based on mathematical algorithms such as, neural network and support vector machine for classification problems. As a matter of fact, these techniques produce satisfactory results, but still suffer from the understandability problem, which are unable to discover understandable rules (i.e. the rules are necessary for the user/expert to understand the problem domain and the classification/prediction tasks). However, this problem can be resolved by integrating between association rule mining and classification.

Associative Classification (AC) approach combines association rules mining and classification tasks. This approach has been shown to be a more accurate classification technique than the traditional methods (Janssens et al., 2003, Antonie et al., 2003, Yin and Han, 2003, Thabtah et al., 2006). Additionally, AC approaches generate rules that are clearer and better for a human expert to understand an application domain (Antonie and Zaïane, 2002).

This provides the ability to optimize (adjust) and update a rule without affecting the full set of rules. In contrast with decision tree technique, any modification process for a rule necessitates a reshaping of the complete tree (decision tree) (Thabtah, 2007). AC model is constructed using the best rules that are learned and generated from training data (Chang Chien and Chen, 2010).

There have been many efforts proposed to build a classifier model based on CARs (Zhang et al., 2009, Chen et al., 2006b, Vo and Le, 2009, Kannan and Bhaskaran, 2010). Practically, association rules mining and AC approaches globally search all rules that satisfy both *minsupp* and *minconf* thresholds. Many of the rules are discovered by AC approaches but cannot be generated by traditional classification techniques such as C4.5 (Liu et al., 1998, Thabtah et al., 2005). Consequently, extraction of a full set of classification association rules contains an important and significant knowledge. As a result, this contributes to building a real prediction model (Janssens et al., 2003).

2.10 Vertical Data Representation

Several association rules mining and AC approaches have employed a traditional horizontal data format based on level-wise in scanning a database to calculate support values of a frequent itemsets as in Apriori approach (Agrawal and Srikant, 1994, Park et al., 1995, Agarwal et al., 2000). Horizontal data format (representation) suffers from multiple data scans in finding frequent itemsets at each level, which causes a high computation time. Previous studies (Zaki and Gouda, 2003, Thabtah et al., 2005, Zaki et al., 1997) have confirmed that a vertical data format is an efficient method of data representation since, it can be assisted in generating candidate itemsets and supports calculation by operating an intersection between Transaction ID (TID) of items in a data. Most of the current AC approaches employ Apriori approach. However, the horizontal data format method is an exhaustive search method in discovering frequent itemsets to form later association rules. These approaches require much time in generating rules (Zaki et al., 1997, Janssens et al., 2003). A few number of approaches have been tackled and utilized the vertical data format (Zaki and Gouda, 2003, Thabtah et al., 2005), which can be achieved by operating an intersection operation between TID of items, whereas the occurrence of each item in a training data is associated as a list. Figure 2.12 shows an example of a vertical data format representation method, the intersection between C and D items is two TID (TID number 1 and 4). Figure 2.12 is explained as follows:

- Database in step 1 is scanned one time to find the items associated with their TID as vertical TID set of candidate itemset C_1 in step 2. The support value is calculated by using the cardinality of the set of TID that is associated with each item, for example, the support value of B is equal 2.

- Candidate itemsets C_1 are moved to the frequent itemsets L_1 in step 3 if their support values are greater than or equal to *minsupp*.
- Frequent itemsets L_1 are joined up with L_1 to generate candidate itemsets C_2 in step 4. Each candidate itemset is associated with their TID. Therefore, the support value for each candidate itemset C_2 is calculated by using the cardinality of the set of TID that is associated with each itemset, instead of scanning database in step 1 (level-wise in scanning a database).
- Candidate itemsets C_2 are moved to the frequent itemsets L_2 in step 5 if their support values are greater than or equal to *minsupp*.
- Frequent itemsets L_2 are joined up with L_2 to generate candidate itemsets C_3 in step 6. Therefore, the support value for each candidate itemset C_3 is calculated by using the cardinality of the set of TID that is associated with each itemset, instead of scanning database in step 1.

Step1

Database	
TID	Items
1	CDE
2	ACE
3	BDE
4	CD
5	BE

Step2

Vertical TID set of Candidate itemset C_1				
A	B	C	D	E
2	3	1	1	1
	5	2	3	2
		4	4	3
				5

Step3

Vertical TID set of Large itemset L_1			
B	C	D	E
3	1	1	1
5	2	3	2
	4	4	3
			5

Step4

Vertical TID set of Candidate itemset C_2					
BC	BD	BE	CD	CE	DE
	3	3	1	1	1
		5	4	2	3

Step5

Vertical TID set of Large itemset L_2			
BE	CD	CE	DE
3	1	1	1
5	4	2	3

Step6

Vertical TID set of Candidate itemset C_3
CDE
1

Figure 2.12 Vertical data format.

An efficient vertical data format was employed in (Thabtah et al., 2005), in order to extract CARs using a single database scanning, instead of applying level-wise (multi-scan). It was proven that the vertical data format is a more efficient method than the horizontal data format. The use of the vertical data format method has saved and reduced a large number of I/O operations (Dunkel and Soparkar, 1999, Zaki and Gouda, 2003). However, with a very large cardinality size of a TID-list, this may lead to a large intersection time (Zaki and Gouda, 2003, Thabtah et al., 2005).

The vertical data format based on TID intersection method used in association rule mining and AC approaches needs to be adapted to treat fuzzy data.

In Chapter 4, the existing vertical data format method for database scanning (Zaki, 2000, Zaki and Gouda, 2003, Thabtah et al., 2005, Zaki et al., 1997) is adapted in order to improve search method efficiency in generating frequent itemsets (frequent termsets) from fuzzy data. The adaptation method is entitled Enhanced Fuzzy Data Representation (EFDR) which is presented and explained further in Section 4.4.

The proposed FACRM prediction model in Chapter 4 utilized EFDR. The EFDR is performed by scanning a fuzzy data one time, and then generating frequent termsets from a previous iteration. A detailed explanation of the EFDR is described in Section 4.4.

2.11 Artificial Neural Network

Artificial Neural Network (ANN) is the most common techniques widely used in many different domains such as classification like pattern recognition and other problem domains (Wong et al., 1997). ANN acts as a biological neural system which can be considered as a human brain in reasoning. It consists of an input layer, hidden layer, output layer, neurons and weights. The neurons are associated with each other via links, and each link is assigned by a numerical weight. Figure 2.13 shows the structure of Multi-Layer Perception (MLP) ANN.

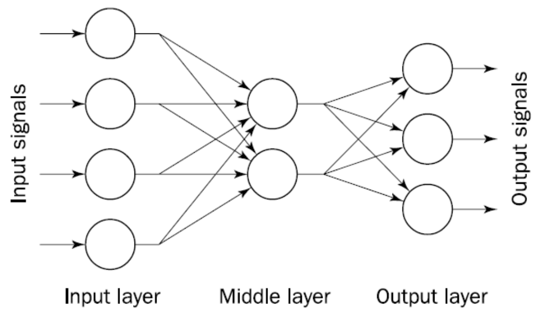


Figure 2.13 The structure of Multi-Layer Perception (MLP) ANN.

Basically, learning in ANN can be accomplished by feeding the data via the input layer, which is then passed to the next layers, hidden and output through the interconnected neurons and activation functions to find solution results for a particular problem. Knowledge is embedded in ANNs based on representing an activation and adjustment weight between neurons (ÖzbakIr et al., 2010).

ANN is one of the most applied models due to its feature in nonlinear processing. The effectiveness of ANN refers to applying a nonlinear function in the hidden layer. However, the architecture of ANN is hard to interpret (Duan et al., 2009), and an experience is required for setting the appropriate neural network model (Abu-Nimeh et al., 2007). Also, it is required to fit an appropriate structure model for each domain (Shin et al., 2005).

The back-propagation neural network is an algorithm commonly used for its characteristics in minimizing the error function. Multilayer Perceptrons (Simon, 1999) are supervised feed-forward networks, which are trained by using the back propagation learning algorithm. Based on training input and output data, the Multilayer Perceptrons model learn how to transform the input data into a particular output. This is frequently used for prediction and classification problems.

There are three main parameters that affect the accuracy of ANN model: (i) number of hidden layers, (ii) number of neurons in each hidden layer and (iii) type of activation functions. Therefore, it is necessary to identify these parameters before constructing the model (network).

This thesis uses ANN model as a comparative technique. The used ANN model is based on a simple Multilayer Perceptrons with a back propagation learning algorithm as implemented in MATLAB 7.6. This has the advantage in reducing user interactions in building ANN model (configuring or setting ANN). ANN is used in Chapter 4 for comparisons using one hidden layer. The number of input neurons and hidden neurons are equal to the number of data attributes. The commonly used activation function for the hidden neurons is tan-sigmoid and the output neurons is linear.

2.12 Support Vector Machine

The Support Vector Machine (SVM) was introduced by Vapnik and others in 1995. SVM is a popular nonlinear data mining technique, used in classification and prediction problems (Sapankevych and Sankar, 2009). The classification in SVM is applied by splitting hyperplane between classes with maximizing the margin between the classes' objects as shown in Figure 2.14 (Abu-Nimeh et al., 2007).

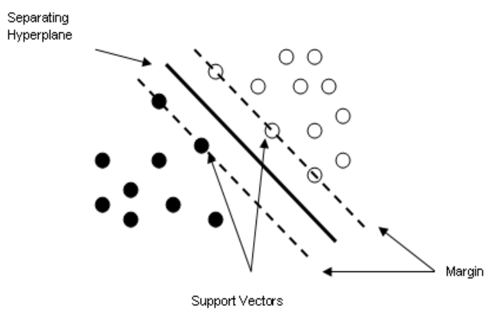


Figure 2.14The classification in SVM.

In spite of the fact that both ANN and SVM are widely applied models and also considered as accurate classification and prediction techniques, they are categorized as non-rule-based techniques (Özbakir et al., 2009). Thus, they are structured as a black-box which is not able to generate high-level rules that can be used to support a human expert in order to understand a problem domain.

In this thesis, the SVM technique with a Radial Basis Function (RBF) kernel is used based on the available platform of LIBSVM (LIBrary for Support Vector Machines) software (Chang and Lin, 2011).

2.13 Stepwise Regression

Stepwise Regression (SR) is one of the commonly used prediction models, its construction is based on an adoption to find a relationship between different independent attributes (input attributes) and a dependent attribute (output attribute). SR uses historical data in order to assess a relationship between a subset of significant independent attributes and a dependent attribute and then build a prediction model. The independent attributes are added to the model until no significant improvement can be made with the dependent attribute.

In SR the initial model is constructed and then the independent attributes are iteratively (systematically) inserted and deleted according to the statistical significance F-test, in order to find a set of independent attributes that have a highest correlation with the dependent attribute. SR aims to select a subset of feature (a subset of input attribute) that maximize the F value. The principle used in SR is to add a variable (attribute), then it is checked for an increasing F value, if the F value decreases, then the attribute is removed from the model (Mendes et al., 2003).

SR has been considered a desirable model, which is used widely as a benchmark prediction models in many studies (Shepperd and Kadoda, 2001, Mendes et al., 2003, Mendes et al., 2007, Azzeh et al., 2011).

2.14 Classification and Regression Trees

Classification and Regression Trees (CART) (Breiman, 1984) is a model that builds a decision tree T in order to predict a dependent (output attribute) y of independent attributes x (Put et al., 2003). CART can be used either for regression if the output is continuous values or for classification if the output is categorical values. It seems that CART is an effective and flexible model in the sense of constructing a nonlinear relationship between the dependent and independent attributes. However, it suffers from data overfitting and generates a large tree which is hard to explain and deal with (Abu-Nimeh et al., 2007).

2.15 Feature Selection Methods

Feature selection (also identified as attribute selection, variable selection, variable subset selection or feature reduction) is a method frequently used in data mining. The purpose of feature selection is to choose new data attributes (feature subset) among the original data, which is an essential part of the pre-processing stage to reduce high dimensionality data (Tan et al., 2006).

Recently, extensive research and investigation has been conducted by researchers in the fields of knowledge discovery and data mining towards improvement of their classification and prediction models in terms of accuracy and performance (Liu and Yu, 2005, Peng et al., 2010, Arauzo-Azofra et al., 2011). The investigations include real-life applications with large and high dimensional databases. Feature selection method is employed to reduce high dimensional data that offers the selection of useful features (feature subset) that are highly predictive regardless of a specific learning algorithm (data mining technique). This kind of feature selection method is called a filter approach. As a result, the learning algorithm can be operated almost better in terms of efficiency and accuracy (Hall, 2000).

The advantages of feature subset selection are (Ding and Peng, 2005):

- Decreasing the predictive computation (improve efficiency) by employing the reduction of data dimension.

- Improving the predictive accuracy by using noise reduction. Noise features are the redundant features that cause increases in prediction error (Manning et al., 2008 , Kumar et al., 2005).
- Producing compact results by selection of representative features that are able to identify a target feature (Hall, 2000). Selection of a small number of features contributes easily to the identification of their relationship with the target feature.

2.16 Evaluation Criteria

A description of the statistical gauges that are used as evaluation criteria to assess the performance of the proposed model is given in this section. The most frequently used benchmark evaluation criteria in different prediction domains are employed in the next chapters for a comparison of the proposed model with other prediction models. The gauges are given by equations 2-8 to 2-15 (Polydoras et al., 1998, Grivas and Chaloulakou, 2006, Sousa et al., 2007, Wang et al., 2008, Duan et al., 2009, Wu and Lo, 2010, Azzeh et al., 2010), respectively:

- Mean Absolute Percentage Error (MAPE)

$$\mathrm{MAPE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\left|\frac{\mathrm{PV_i} - \mathrm{RV_i}}{\mathrm{RV_i}}\right| * 100\right) \tag{2-8}$$

- Median Absolute Percentage Error (MdAPE)

$$\mathrm{MdAPE} = \mathrm{median}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\left|\frac{\mathrm{PV_i} - \mathrm{RV_i}}{\mathrm{RV_i}}\right| * 100\right) \tag{2-9}$$

- Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \left(|\mathrm{PV_i} - \mathrm{RV_i}|\right)$$

(2-10)

- Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{PV_i} - \mathrm{RV_i})^2$$

(2-11)

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{PV_i} - \mathrm{RV_i})^2}$$

(2-12)

- Normalized Mean Absolute Error (NMAE) (Han et al., 2003)

$$\mathrm{NMAE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{PV_i} - \mathrm{RV_i}|}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{RV_i}} * 100\%$$

(2-13)

- Normalized Root Mean Squared Error (NRMSE) (Han et al., 2003)

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{PV_i} - \mathrm{RV_i})^2}}{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{RV_i}} * 100\%$$

(2-14)

- Pearson Product-Moment Correlation Coefficient r (Pearson r) (Quek et al., 2006)

$$\mathrm{r} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{RV_i} - \overline{\mathrm{RV}})(\mathrm{PV_i} - \overline{\mathrm{PV}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{RV_i} - \overline{\mathrm{RV}})^2 \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{PV_i} - \overline{\mathrm{PV}})^2}}$$

(2-15)

where, PV: the predicted output value, RV: the real output value, N: Total number of the comparison records, $\overline{\mathrm{RV}}$: The mean of real output, $\overline{\mathrm{PV}}$: The mean of predicted output, $\mathrm{Uncorr{:}Uncorrelated} = 1 - \mathrm{r}$.

The above evaluation measures have been used widely to assess prediction models. The Pearson r is used to measure the relationship between two variables (the predicted and real output value). The value of Pearson r has a range from -1 to +1; when r equals 0, the predicted and real output values are uncorrelated; when r equals 1, the predicted and real output values are approximately the same, and when r equals -1, the predicted and real output values are approximately the same with opposite direction. In view of the fact that the MAPE is sensitive to some extreme error values, MdAPE is applied which is less affected by the outliers' error values than the MAPE.

MAE measures the arithmetic average value (magnitude) of all prediction errors, i.e. summing the absolute error values (absolute values of the difference between real and predicted values). MAE is considered as the most natural measure of the average error. The calculation of RMSE is illustrated in three steps. Firstly, the squared error values (the squared difference between the real and predicted values) is found. Secondly, the total square error (the summation of the squared error values) is divided by the number of error values, which produces MSE. Finally, the square root of MSE is calculated to find RMSE. In RMSE, each error affects the total error values in proportion to its squared errors. Large error values have a relatively high effect on the total square error compared to the effect of the smaller error values. It should be noted that, RMSE is always greater or equal to MAE. RMSE detects the error values variations in sampling errors (error values). The higher the RMSE, the larger the variations and vice versa. If RMSE is closer to MAE, that means the model is consistent (Willmott and Matsuura, 2005, Witten and Frank, 2005).

Since all absolute percentage error values for all models are not normally distributed, as examined by one-sample Kolmogorov-Smirnov (K-S) test, thus a nonparametric test, in particular the Wilcoxon rank sum test, is used to examine the statistical significance between the proposed FACRM model and other prediction models in Chapter 4.

2.17 Summary and Conclusions

In this chapter a general overview and discussion of the main related themes were reviewed. A brief introduction and background of the most important topics concerning knowledge discovery and data mining, data mining tasks, association rules mining and fuzzy logic were highlighted. A range of the existing works in the field of fuzzy association rules was presented where the fuzzy inference system and the similarity that applied in the association rules were demonstrated. Associative classification approaches, vertical data representation, Artificial Neural Network (ANN), Support Vector Machine (SVM), Stepwise Regression (SR), Classification and Regression Tree (CART) and feature selection method were reviewed and the evaluation measures that will be used in the experiments to evaluate and validate the proposed models in Chapters 3, 4, and 5 were identified and described.

In summary the main research issues discussed in the literature which have been motivation factors for the research reported in this thesis are:

- The extraction of association rules from a quantitative data.
- The limitations of using single *minsupp*.
- The lack of focus on a post processing method and a diverse knowledge (representative rules).
- The limitations of the current AC approaches in generating dominating rules (significant rules).

- The necessity for a reliable prediction model that is applicable to different data sets of different application domains.
- The effect of using high data dimension (all data attributes (features)) on a prediction model.

In this thesis, the main focus is to explore the research issues highlighted above by investigating approaches and proposing a prediction model for building a knowledge base with useful knowledge to provide an effective and reliable prediction. Furthermore, the idea behind proposing a feature selection method is to improve the performance of the prediction model.

In summary, the research issues that are addressed through this research in order to build a reliable prediction model are as follows: (i) the problem of extracting association rules from quantitative attributes, which is tackled by applying fuzzy clustering techniques to transform quantitative data into fuzzy ones, (ii) the use of single *minsupp* for whole database that causes the rare item problem, which is attempted by employing of multiple support thresholds approach and diversification method, (iii) the current Associative Classification (AC) approaches suffer from the following limitations: generating high number of rules, using single *minsupp*, using an objective measure (*minconf*), applying level-wise like Apriori fashion, generating non-dominating (insignificant) rules and the process of rules extraction. These limitations are undertaken by employing the recently improved multiple support approach, adapting vertical data format scan and applying diversification method, (iv) high dimensional data is the one of the most important issues that affects on a prediction model error and performance. This issue is addressed by developing a feature selection method entitled Weighting Feature Selection (WFS).

CHAPTER THREE

3 PREDICTION MODELS BASED ON FUZZY ASSOCIATION RULES MINING

3.1 Chapter Overview

This chapter aims to provide an insight into two Knowledge Discovery (KD) models, which are developed to extract knowledge that can be applied to predict a future value. The first model integrates Fuzzy C-Means (FCM) and Apriori approach and is applied for road traffic performance prediction. The FCM is used to define the membership functions of fuzzy sets and Apriori approach is employed to identify the Fuzzy Associations Rules (FARs). The proposed model extracts knowledge from a database for a Fuzzy Inference System (FIS). The knowledge extraction process and the model performance are demonstrated through two case studies of the road traffic data set with different sizes. The experimental results show the capability of the proposed KD model in FARs based knowledge extraction. The second model proposes an approach, called Diverse Fuzzy Rule Base (DFRB), to extract the FARs which are used later to build a prediction model for predicting a future value. This approach also aims to ensure high quality and diversity of the FARs. This is achieved through four phases. Firstly, the integration of FCM and MSapriori approach is employed to extract the FARs. The second phase calculates the correlation values for these FARs, and performs an efficient orientation for filtering FARs as a post-processing method. In the third phase, the FARs diversity is maintained through the clustering of FARs, based on the concept of the sharing function technique used in multi-objectives optimization. Fourth phase, these FARs provide the knowledge base to be utilized within the FIS for a reliability prediction and evaluation. Reliability refers to the trade-off between minimizing prediction error and ensuring rules diversity. Experimental results for two case studies have shown that the second model of DFRB approach predicted the future values effectively for a wide range of the input data sets, thus, outperforming the first model and the model reported in the literature. The results also demonstrate the merits and effectiveness of the proposed approach in building a reliable prediction model.

The rest of this chapter includes: a review of the related work in section two; a discussion of the case studies in section three; details of the proposed first prediction model in section four; an analysis of the second proposed prediction model in section five; and finally extracted conclusions.

3.2 Introduction and Related Work

Prediction is a vital and important task in Data Mining (DM). The aim of the proposed prediction model is to predict a future value accurately. This can be achieved by building a model that generates and evaluates a set of rules for prediction. Association rules mining is one of the most important tasks in DM research and is increasingly attracting the attention of researchers. Most of the common association rules algorithms are based on level-wise, such as Apriori (Agrawal and Srikant, 1994), and others use the tree structure namely Frequent Pattern Growth (FP-Growth) (Han et al., 2000).

A fuzzy approach is widely used in intelligent systems, since it is very simple and similar to the human way of thinking. Fuzzy Logic is defined as a knowledge representation from data using a set of mathematical theories based on membership functions. The fuzzy approach can be used to assist in extracting knowledge from a database by transforming quantitative data (crisp data) into fuzzy data. Thus the process is achieved through the identification of the membership functions and Fuzzy Association Rules (FARs), as well as the application of proper Knowledge Discovery (KD) and Data Mining (DM) techniques. As a result, a fuzzy clustering technique is applied to handle the problem of a quantitative data. It can effectively convert a quantitative data into fuzzy data by finding fuzzy sets (fuzzy terms or linguistic terms).

Fuzzy association rules mining has been successfully applied to a wide range of classification and prediction applications. Huang et al. (Huang et al., 2006) proposed a fuzzy data mining approach to discover rules by applying an Apriori approach entitled “modification of the fuzzy transaction data-mining algorithm” while adapting the discovered rules for the training Adaptive Network based on Fuzzy Inference System (ANFIS). The approach is applied in the human resources department for predicting future employee performance, in either suitable projects or positions. This approach (Huang et al., 2006) was tested for a small data set with some noise. The member functions of fuzzy set are defined and known in advance. However, it could be adapted for a small data set, but it is not feasible in the case of a large data set. Lu et al. (Lu et al., 2003a) compared two approaches for one output prediction value, applied in a quantitative data set taken from the University of California, Irvine (UCI) machine learning Repository called Abalone. In the first method, FCM and Apriori approach were used for extracting FARs as well as Genetic Algorithm (GA) which was applied for tuning the fuzzy sets. The second method proceeded as the first method, but it used variable thresholds in the prediction. The difference in accuracy prediction between the two approaches was too small. Zhang and co-authors (Zhang et al., 2005) implemented fuzzy mining algorithm to find out the implicit knowledge (rules) from Iris numeric data set. They concluded that when a suitable *minsupp* and *minconf* are selected, valuable

FARs can be obtained. Hong et al. (Hong et al., 2004) also proposed an approach for mining FARs, which is based on AprioriTid algorithm to extract FARs from quantitative supermarket purchase data. It used the maximum cardinality with the highest summation value of the fuzzy set (linguistic term) among the fuzzy sets of each data set attribute. Hence the number of fuzzy sets is similar to the original data set attributes as far as the latter mining processes are concerned. In this case one fuzzy set is not enough to reflect the origin data set attribute.

Several algorithms are proposed for generating association rules which have support and confidence values higher than user-specified thresholds (Kryszkiewicz, 1998). Similarly, there are several techniques that can be conducted to prune the huge number of such association rules and transform them into more representative ones (Kryszkiewicz, 1998, Kryszkiewicz, 2009, Kryszkiewicz and Rybinski, 1999). Marzena (Kryszkiewicz, 1998) introduced an approach to obtain representative rules from a large set of association rules using cover operator, where these rules are based on satisfying the *minsupp* and *minconf* measures. The use of such measures to generate these rules could be affected at the representative rules level. The representative rules includes small numbers of rules, which decrease the accuracy when they are evaluated and validated. Many techniques were offered by scholars suggesting objective measures to evaluate the association rules (Le et al., 2009, Suzuki, 2009). Lenca et al. (Lenca et al., 2008) proposed an approach for selecting the most interesting rules based on improvement of the objective measurements. Their approach used a Multi-Criteria Decision Aid (MCDA) method to sustain these rules for a non-expert user in a specific domain.

In this chapter, two prediction models are proposed. The proposed models are implemented, tested and verified through a set of experiments and then compared with the existing work to demonstrate their merits and capabilities. The first model is based on the integration of Fuzzy C-Means (FCM) and Apriori approach for extracting FARs, and applied in road traffic domain for the prediction of a future value. The aim of the first model is to build a first stage of the proposed Knowledge Discovery (KD) model for prediction, by employing the well-known and popular Apriori approach. The first stage is considered an initial prediction model (prototype model), therefore, the importance of association rules mining is depicted and the limitations of the existing approach are identified practically. The second model is based on the proposed approach called Diverse Fuzzy Rule Base (DFRB), which aims to extract a robust (best) and diverse fuzzy rule base that enables the prediction of future values effectively. These fuzzy rules base are generated from FARs, and should be sustained and proved. The proposed approach facilitates the trade-off between a prediction error and the diversity within FARs. The second model basically enhances the first model to capture the rare termsets related rules, and is applied to the same road traffic domain for prediction of the future value and other benchmark data set called Abalone.

3.3 Case Studies

Before presenting the two prediction models, this section describes the details of the case studies used to demonstrate the capability and perform a comparative study of the proposed models.

3.3.1 Road Traffic Case Studies

Two data sets related to the road traffic problem have been employed for the model performance analysis in Section 3.4. Traffic state prediction (including traffic flow (traffic density) and traffic demand) has long been regarded as a critical concern for intelligent road traffic systems. The road traffic data has been generated using a traffic simulation model, (called the METANET macroscopic flow model) (Messmer, 2007). Each record consists of:

- Traffic state, which is represented by: traffic demands in road 1 (the numbers of vehicles that need to use the road 1), traffic demands in road 2, traffic density in road 1 (the number of vehicles that are using road 1 per km), and traffic density in road 2. Figure 3.1 shows information about the input of road traffic data set.
- Predicted Average Travel Time (ATT) (ATT is the total average time required for a vehicle to cross the traffic network).

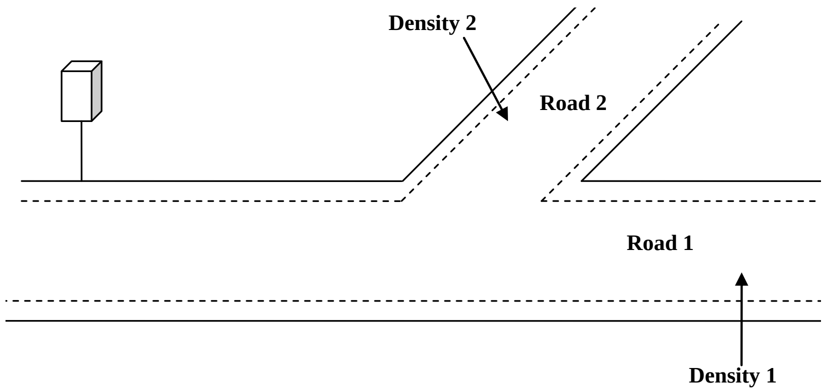


Figure 3.1 Information of the input road traffic data set.

a) Small data set size

This data set of 100 records contains the following fields (Demand 1, Demand 2, Density 1, and Density 2 as input, and ATT as output) as shown in Figure 3.1 and Table 3-1, and the statistical information regarding this data set is shown in Table 3-2. Figure 3.2 and Figure 3.3 show the analysis of the data set, to clarify the distribution and consistency. The data set is divided into 75 records for training and 25 records for testing.

The boxplot (box-and-whisker diagram) provides a suitable method of graphically describing groups of numerical data, and is used for detecting any outlier (noisy) data. From these two figures the data set was found to be consistent and without any outlier data. Also all fields were distributed in different ranges of value for this reason the data fields (attributes) are divided into two figures, but our aim is to find out if there are any outlier cases for each field separately.

Table 3-1 Part of the road traffic data set.

Case No.	Demand 1	Demand 2	Density 1	Density 2	ATT
1	6030	316	5	87	629.7
2	1147	1638	62	50	414.4
3	1277	797	10	14	233.5
4	4061	1198	38	75	581.2
5	2382	487	30	86	530
6	2416	2306	88	99	723.1
7	899	552	60	28	331.1
8	687	1669	81	61	509.4
9	739	3049	60	58	494
10	3504	431	32	72	502.2
:	:	:	:	:	:

Table 3-2 Statistical information of the road traffic data set.

	Demand 1	Demand 2	Density 1	Density 2	ATT
minimum	185	123	1.000	1.000	222.800
maximum	6986	3498	99.000	99.000	729.800
mean	3670.110	1640.450	48.850	50.320	515.715
Standard deviation	2060.326	919.690	28.919	27.376	109.980
Correlation	0.438	0.257	0.398	0.527	1.000

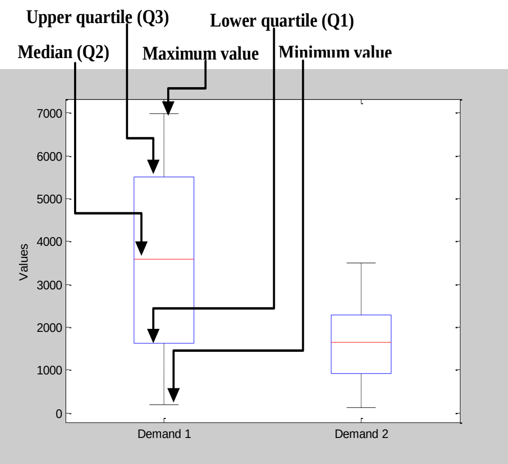


Figure 3.2 Analysis of the road traffic data set for Demand 1 and Demand 2.

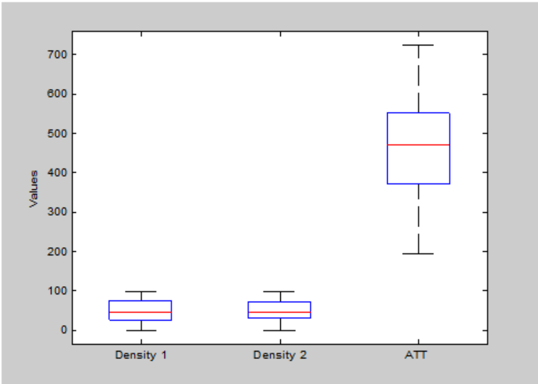


Figure 3.3 Analysis of the road traffic data set for Density 1, Density 2, and ATT.

b) Large data set

The second road traffic data set used has a large size of 1,000 records with the following fields (Demand 1, Demand 2, Density 1, and Density 2 as input, and ATT as output). The data set is divided into 750 records for training and 250 records for testing.

The boxplot in this data set was used to ensure its consistency as done in the previous case.

3.3.2 Abalone Benchmark Data

The Abalone data taken from University of California, Irvine (UCI) of machine learning repository (Frank and Asuncion, 2010) was employed to conduct a comparative study of the proposed model in Section 3.5 with other model reported in the literature. Each record in the data set consists of:

- Abalone body description, which is represented by: Sex (M, F, and I (Infant)), Length (Longest shell measurement), Diameter (perpendicular to length), Height (with meat in shell), Whole weight (whole Abalone), Shucked weight (weight of meat), Viscera weight (gut weight (after bleeding)), and Shell Weight (after being dried).
- Abalone rings, which is used to predict the Abalone age.

Abalone data set is divided into 3,133 records for training and 1,044 records for testing. Note that the Abalone data set is used only to validate the second model. The first model has already been validated in the literature using this data set (Lu et al., 2003a).

3.4 The Proposed First Prediction Model

3.4.1 The Model Description

The first proposed KD model extracts fuzzy rules for building a KB from database, and is based on the work of Huang et al. (Huang et al., 2006), and Lu et al. (Lu et al., 2003a). The KD model utilizes the following two methods:

- FCM is used as an automatic system to transform the quantitative data set into fuzzy sets (terms).

- Apriori approach is used for extracting fuzzy termsets (frequent itemsets) from fuzzy data set based on interesting measures (*minsupp* and *minconf*). Throughout the rest of the thesis, the term itemsets corresponds to its termsets.

Figure 3.4 shows the model steps: (i) getting the data set from the database, which is analyzed for consistency and any noisy data set will be removed, (ii) transforming the quantitative data set into fuzzy sets using FCM, (iii) applying the Apriori approach to extract FARs, and then saving these rules in Knowledge Base (KB), (iv) using Fuzzy Inference System (FIS) to command the KB for a prediction and (v) testing the feasibility of the KD model in the road traffic case studies.

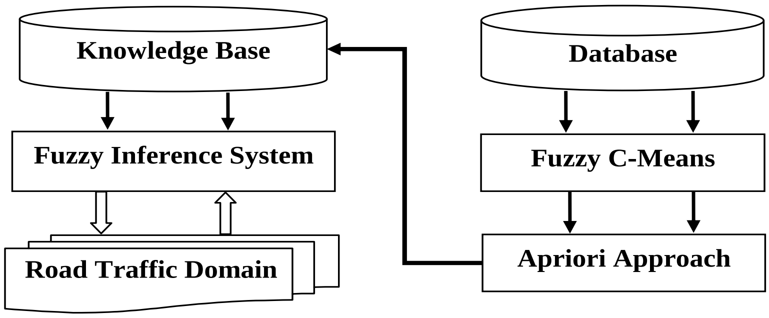


Figure 3.4 The proposed first KD model steps.

The following definition (notation) is used in KD model:

- Field: Attribute (item or column) of the crisp input data.
- Record (Case): Row with all fields.
- Term: Fuzzy set class (fuzzy term).
- x_{ij}: Value of the crisp input data.
- $\mu(x)_{if}$: Fuzzy set value.
- Sum_{jf}: Summation of each fuzzy term for all records.
- Termset: A set of terms contains one term or more.
- C_k: Contains candidate termsets, $1 \leq k \leq n$, n = maximum number of the fields.

- L_k: Contains large termsets, $1 \le k \le n$, n = maximum number of the fields.
- *minsupp*: Minimum support threshold value (observing that *minsupp* =1.875 for the small data size, and *minsupp* =4 for the large data size).
- *minconf*: Minimum confidence threshold value (observing that *minconf* =0.9 for the small data size, and *minconf* =0.8 for the large data size. These value are selected based on many experiments run to find out the appropriate ones that able us to extract useful rules. Thus the error is minimized).

The proposed KD model as shown in Figure 3.5 works as follow:

1. FCM is used to cluster the data into terms and then to determine the centre of each fuzzy set. In addition, the maximum and minimum values for each field of input data set are found out.
2. Data set is converted into a fuzzy data set, using one of the standard membership functions (the triangular and trapezoid membership functions (Hong et al., 2004)).
3. Support value is calculated for each term by summing the fuzzy values in each term for all records using Equation 3-1, then this summation value is stored in the candidate termset C_1.

$$\mathrm{Sum}_{\mathrm{jf}} = \sum_{i=1}^{n} \mu(x)_{\mathrm{if}} \tag{3-1}$$

4. Terms are moved to L_1, which are greater than or equal to *minsupp*.
5. Terms are joined up and combined, as (L_1 join L_1) = $\{\{c[1], c[i]\}, \{c[1], c[i+1]\} \ldots \{c[1], c[n]\}\}$, where c[1]: the first fuzzy term, c[i]: the second fuzzy term and $c[n]$: the last fuzzy term, where $c[1] \cap c[i] = \emptyset$, $c[1] \cap c[i+1] = \emptyset \ldots c[1] \cap c[n] = \emptyset$ (i.e, the terms for each termset do not belong to the same field). Once every termset is

stored in the candidate termset C_2, the support value for each termset will be calculated using a minimum operator for the fuzzy values between the terms in the termset. In addition, the result of the minimum values in that termset is summed for all records. Finally, the results' summations will be stored in the candidate termset C_2.

```
Begin
FCM; {clustering data set}
Find the fuzzy sets of the quantitative data set, based on FCM.
Calculate the summation of the membership value for each fuzzy term with all records using
Equation 3-1.
      IF Sum_jf ≥ minsup Then
      Insert the fuzzy term into L_1, L_1={frequent termsets}
For (k = 2; L_{k−1} ≠ ∅; k++) do
     C_k = generate candidate from L_{k−1}(join L_{k−1} called (p) with L_{k−1} called (q));
         {
           Insert into C_k
           Select termset; p.term_1, p.term_2 ... p.term_{k−1}, q.term_{k−1}
           From p, q
           Where p.term_1 = q.term_1 ... p.term_{k−2} = q.term_{k−2} , p.term_{k−1} ≠ q.term_{k−1}
         }
     For each termset c ∈ C_k do
         Check the all sub-termsets of all termsets in C_k , and it should be a frequent termsets
         in L_{k−1}
         For each (k − 1) subset s of c do
             IF s ∉ L_{k−1} Then
                 Delete c from C_k
             EndIF
         EndFor
     EndFor
     For each terrmset candidate in C_k do
         Calculate the support value using Equation 3-1.
         IF Sum_jf ≥ minsup Then
             Insert the fuzzy termset into L_k, L_k={frequent termsets}
     EndFor
EndFor
Select the frequent temsets including the target attribute (output attribute).
Form the frequent termsets (rules) that exist in L_2 to L_k under the form "IF-Then".
For each rule
    Calculate the confidence value for each rule using Equation 3-2.
    IF CV ≥ minconf Then
       Accept the rule.
EndFor
Check the rules for contradiction.
Insert all the accepted rules in KB.
Infer the existed rules in KB using FIS.
EndBegin
```

Figure 3.5 The proposed first KD model.

6. Termsets greater than *minsupp* are moved to L_2.

7. Termsets are joined up and combined again as $L_2 = \mathrm{p}$ join $L_2 = \mathrm{q}$, where $\mathrm{p.term}_1 = \mathrm{q.term}_1 \ldots \mathrm{p.term}_{k-2} = \mathrm{q.term}_{k-2}, \mathrm{p.term}_{k-1} \neq \mathrm{q.term}_{k-1}$. This combination is based on every sub-termset of the candidate termset that exists in C_k which should be frequent termset in the previous large termset of L_{k-1}. Also the terms for each termset in C_k do not belong to the same field.
8. Termsets are stored in the candidate termset C_3, then support value is calculated for each candidate termset.
9. Termsets and their support values in C_3 greater than or equal to *minsupp* are moved to L_3.
10. Termsets are joined up and combined, until L_n is empty.
11. Termsets are pruned by selection of the termsets including the target attribute. As a consequence, termsets are formed as IF-Then form, then the Confidence Value (CV) is calculated based on Equation 3-2. The rules with (CV) greater than or equal to *minconf* are accepted. The contradiction rules are removed.

$$\mathrm{CV} = \frac{\sum[(\mathrm{IF}) \cap (\mathrm{Then})]}{\sum(\min(\mathrm{IF}))} \tag{3-2}$$

The extracted rules are stored in KB, which will be used later in the FIS.

3.4.2 *Experimental Results and Analysis*

For analysis and validation purposes, the methodology discussed in the previous section is applied in road traffic control management. Two case studies with different data set sizes of road traffic are considered for predicting the Average Total Time (ATT) of the traffic.

3.4.2.1 *Example: How the Proposed First KD Model Works*

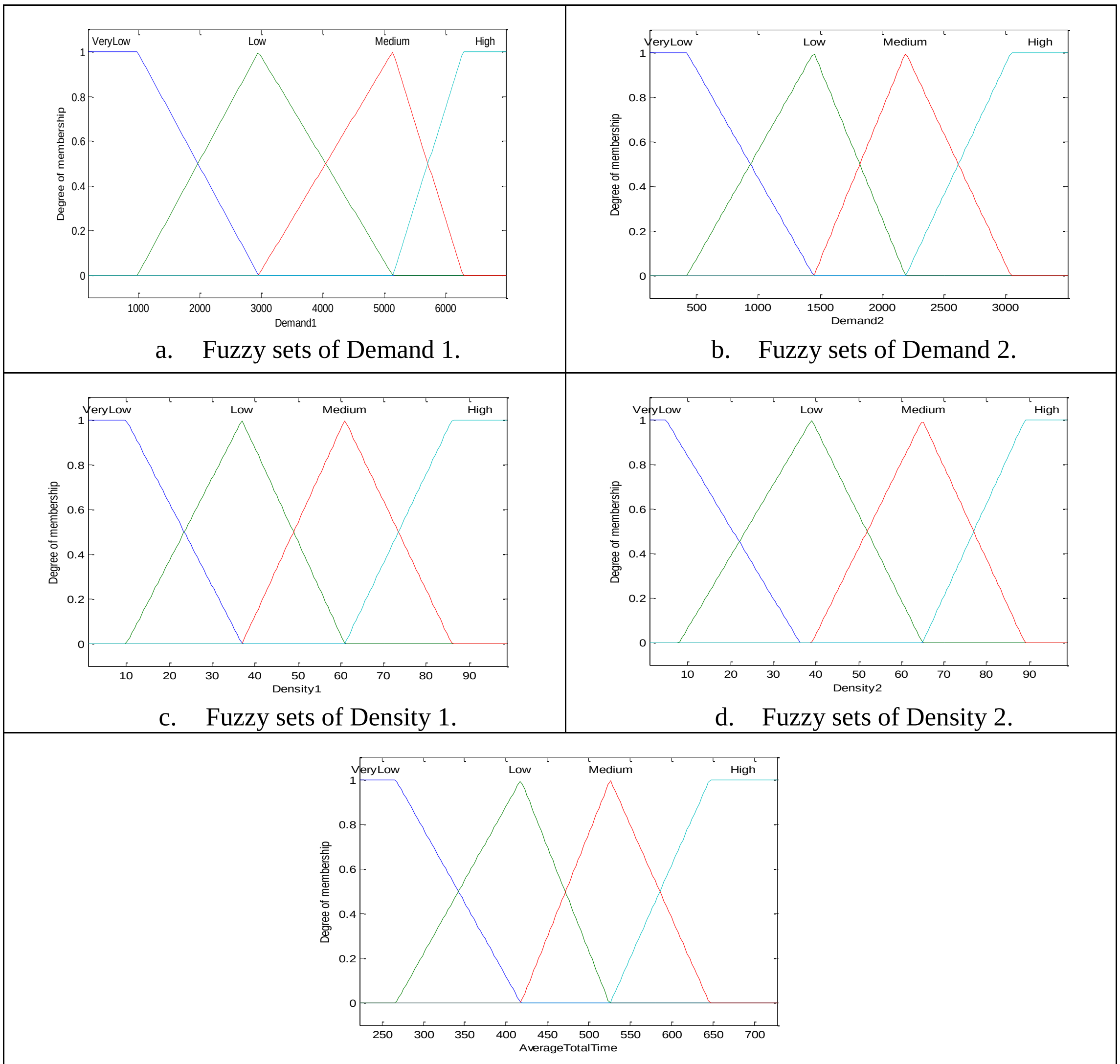
This example illustrates the steps of the KD model that is applied in the small and large size of road traffic data set in order to test the feasibility. These include:

1. Table 3-3 shows the minimum, maximum and centre values of the crisp input data. This helps to determine the fuzzy sets by using these values to find the parameters of the membership functions employing FCM.

Table 3-3 Centers of the data set with minimum and maximum values for each field.

	Demand 1	Demand 2	Density 1	Density 2	ATT
Minimum	185	123	1	1	223
Maximum	6986	3498	99	99	729
Centre 1	990	426	10	8	276
Centre 2	2960	1449	37	39	411
Centre 3	5143	2191	60	65	547
Centre 4	6291	3041	86	89	675

2. Figure 3.6 represents each field and its membership functions (Note: all fields have four fuzzy classes including: Very Low (VL), Low (L), Medium (M) and High (H). For abbreviation, each fuzzy class (fuzzy set) is mapped into numbers, for example, Demand1.VL→1, Demand1.L→2...ATT.H→20). The aim of Figure 3.6 is to illustrate the fuzzy sets of each field for the road traffic of the small size after using FCM. Table 3-4 shows the part of the fuzzy data set, which shows that each value x_{ij} in the original data set belongs to two terms $\mu(x)_{if}$ with different membership values.



a. Fuzzy sets of Demand 1.
b. Fuzzy sets of Demand 2.
c. Fuzzy sets of Density 1.
d. Fuzzy sets of Density 2.
e. Fuzzy sets of Average Total Time.

Figure 3.6 The membership functions for each field used in this case study. (a) the fuzzy sets of Demand 1. (b) the fuzzy sets of Demand 2. (c) the fuzzy sets of Density 1. (d) the fuzzy sets of Denisty 2. (e) the fuzzy sets Average Total Time.

Table 3-4 Part of fuzzy data set for the road traffic.

Demand1				Demand2				Density1				:
VL (1)	L (2)	M (3)	H (4)	VL (5)	L (6)	M (7)	H (8)	VL (9)	L (10)	M (11)	H (12)	:
0	0	0.23	0.77	1	0	0	0	1	0	0	0	:
0.92	0.08	0	0	0	0.75	0.25	0	0	0	0.96	0.04	:
0.85	0.15	0	0	0.64	0.36	0	0	1	0	0	0	:
0	0.50	0.50	0	0.25	0.75	0	0	0	0.96	0.04	0	:
0.29	0.71	0	0	0.94	0.06	0	0	0.26	0.74	0	0	:
:	:	:	:	:	:	:	:	:	:	:	:	:

3. Table 3-5 shows part of the candidate termset of C_1 after the calculation of support value for each term. For instance, the summation value for the term (VL) in (Demand 1) in the small size of road traffic data set is equal to (26.6).

4. Table 3-6 shows part of the large termset of L_1 that contains the terms whose support values are greater than or equal to *minsupp*.

Table 3-5 Part of C_1.

Term	Support
{1}	26.6
{2}	27.2
{3}	24
:	:

Table 3-6 Part of L_1.

Termet
{1}
{2}
{3}
:

5. Table 3-7 shows part of the support value calculation, whereas Table 3-8 shows part of the candidate termsets of C_2.

Table 3-7 Support value calculation.

1	5	Support
0	1	0
0.93	0	0
0.85	0.64	0.64
:	:	:
		$\Sigma = 11.08$

6. Table 3-9 shows the large termset of L_2, with support value ≥ *minsupp*.

Table 3-8 part of C_2.

Termset	Support
{1,5}	11.08
{2,7}	10.46
{2,10}	28.43
{2,17}	22.08
{7,10}	4.79
{7,13}	3.9
{12,14}	10.58
{12,24}	10.13

Table 3-9 part of L_2.

Termset
{1,5}
{2,7}
{2,10}
{2,17}
{7,10}
{12,14}
{12,24}

7. Each sub-termset of the candidate termset exists in C_k and should be considered as frequent termset in the previous large termset of L_{k-1}. For example, these sub-termset={{2,7},{2,10},{7,10}} of the termset ={2,7,10} of C_3 in Table 3-10 should have existed in the previous large termset of L_2 in Table 3-9.
8. Table 3-10 shows part of the candidate termsets that are stored in C_3, and then the support value is calculated and stored in C_3, as shown in Table 3-10.
9. Table 3-11 shows part of the large termset of L_3.
10. Combine termsets, until L_n is empty.

Table 3-10 Part of C_3.

Termset	Support
{2,7,10}	4
{2,7,17}	7.51
{12,14,24}	7.32

Table 3-11 Part of L_3.

Termset
{2,7,17}
{12,14,24}

11. Select the frequent termsets in L_2 to L_n that includes the target attribute (i.e. ATT). Alongside, form the selected termsets under "IF-Then" forms (for example, form the frequent termset of {2, 17} as IF 2 Then 17), then CV is calculated based on the Equation 3-2, and it accepts the rules that are greater than or equal to *minconf* as it is shown in Table 3-12. Then, the contradicting rules are removed, based on the CV. If CV of the rules are the same, then the support value is used. For example, if two rules are found as follows:

IF Demand 1=High and Density 1=Medium Then ATT=Medium, $CV = 0.72$,
IF Demand 1=High and Density 1=Medium Then ATT=High, $CV = 0.61$.

The first rule is selected because it has a higher CV than the CV of the second one.

Table 3-12 Part of CVcalculating.

Frequent termset	(IF-Then) form	Σ[min(IF)]	Σ[(IF) $\cap$ (Then)]	CV
{2,17}	IF 2 Then 17	26.61	22.08	0.83
{12,14,24}	IF 12, 14 Then 24	10.58	7.32	0.69

3.4.2.2 Results Analysis

Figure 3.7 shows a graph of the real ATT (output) values and the predicted ATT (output) of the proposed first KD model for the testing data of the small data set. The MAPE has been calculated using Equation 2-8, which is equal to 9.11 % at *minsupp* and *minconf* equal to 1.875 and 0.9 respectively. The graph shows that the difference between the two plotted lines is relatively small, because a consistent training data (without noisy data), and FCM are used and the appropriate *minsupp* and *minconf* are selected.

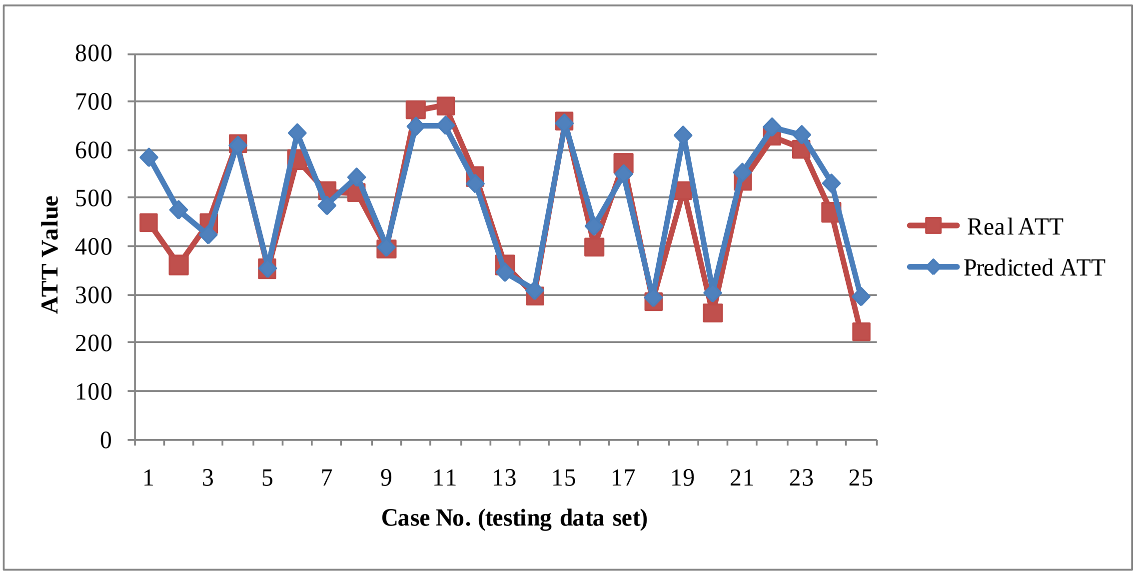


Figure 3.7 The difference between real and predicted ATT.

Figure 3.8 depicts the sensitivity analysis of *minsupp* and *minconf* values on the MAPE for the small data set. The graph of *minconf* 0.9 at *minsupp* 1.875 shows the minimum MAPE value of 9.11%, and it contains rules that cover most cases. *minconf* less than 0.9 will increase the MAPE, this is explained by producing a large number of rules (the decrease in *minconf* implies an increase in the deviated rules, and causes noise for the FIS). It is noted that *minconf* greater than 0.9 will also lead to an increase in the MAPE. Again this is explained by producing a small number of rules, which does not give robust results for the FIS (the increase in *minconf* implies a decrease in the number of relevant rules). The graph of *minconf* 1 is more affected by *minsupp* than the others, in other words, *minsupp* has a large influence on high *minconf* values. The graph in Figure 3.8 demonstrates that the prediction error is sensitive to *minsupp* and *minconf* values. The selection of an appropriate number of rules for accurate prediction depends on the selection of the *minsupp* and *minconf* values.

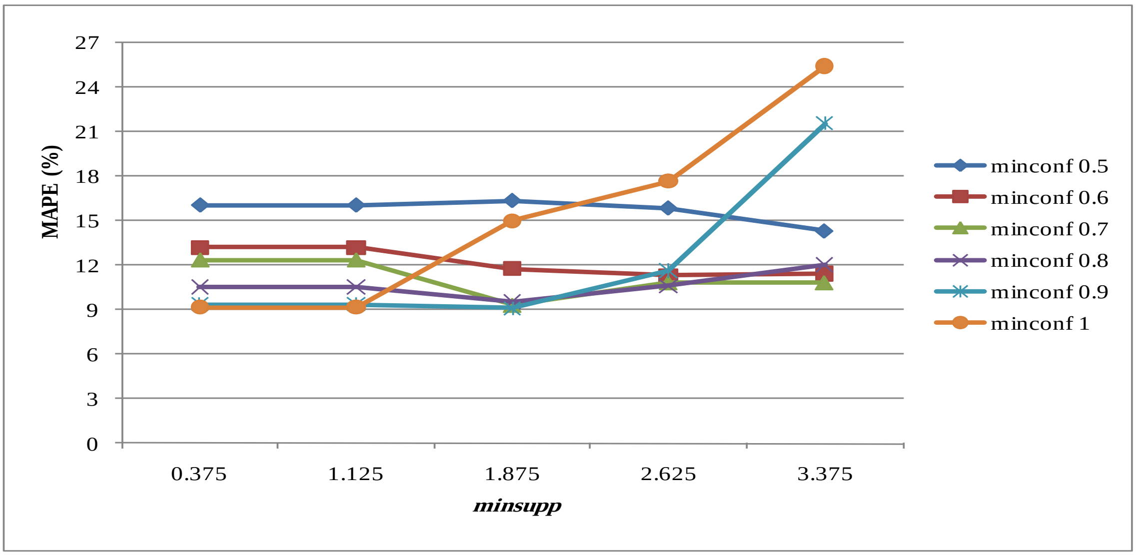


Figure 3.8 The MAPE over different *minsupp* and *minconf*.

Figure 3.9 shows that the graph of *minsupp* equal to 1.875 at *minconf* value 0.9 gives the minimum MAPE, because it contains rules that cover most cases. Also, the figure shows that using *minsupp* less than 1.875 will increase the MAPE, which is explained by producing a large number of rules (decreased *minconf* leads to an increase in the number of rules, and makes it noisy for the FIS). On the other hand, the figure shows that using *minsupp* greater than 1.875 will increases the MAPE, this is explained by producing a small number of rules, which does not give robust results for the FIS (the increase in *minconf* leads to a decrease in the number of rules (does not cover most cases)).

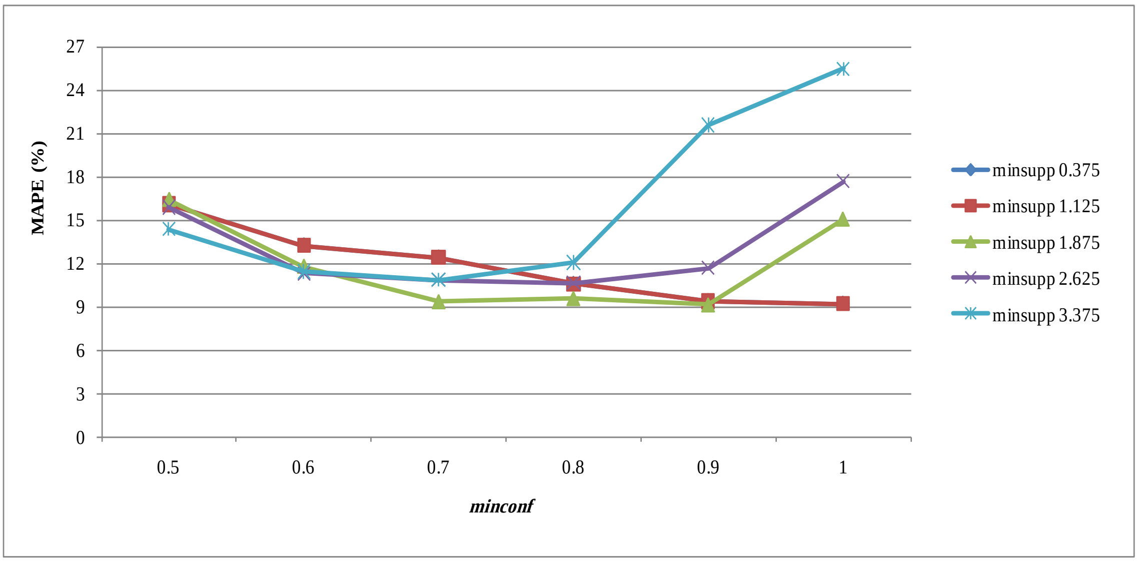


Figure 3.9 The MAPE over different *minconf* and *minsupp*.

Figure 3.10 shows the similar sensitivity analysis results for the large data set. The graph of *minconf* 0.8 at *minsupp* 4 gives the minimum MAPE of 8.5 %. The MAPE value of the large data set size is less than the value of the small data size. This can be explained by the use of a consistent and cooperative large training data set. The graph shows that using *minconf* less than 0.8 will increase the MAPE. This is explained by producing a large number of rules, some of which are irrelevant. The *minconf* greater than 0.8 increases the MAPE. Again this is explained by producing a small number of rules, which does not give robust results for the FIS.

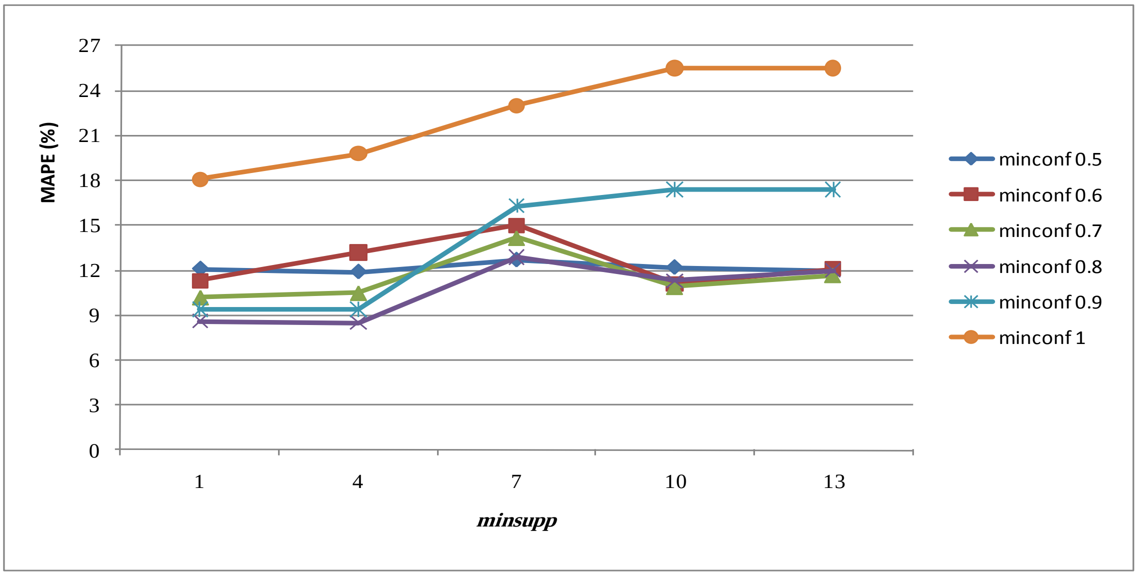


Figure 3.10 The MAPE over different *minsupp* and *minconf*.

Figure 3.11 shows all graphs of the *minconf* over the *minsupp* of both small and large data sets, the values of 0.375-3.375 represent the *minsupp* of the small size data set, whereas the values of 1-13 represent the *minsupp* of the large size data set. Figure 3.11 indicates that using a large data set in the training process will reduce the MAPE. It is worth noting that an increase in the training data size implies a decrease in the error, since it produces rules that cover most of the cases. From Figure 3.11, it can be seen that using different data set sizes and types require the usage of different *minsupp* and *minconf* for producing quality rules to minimize the error.

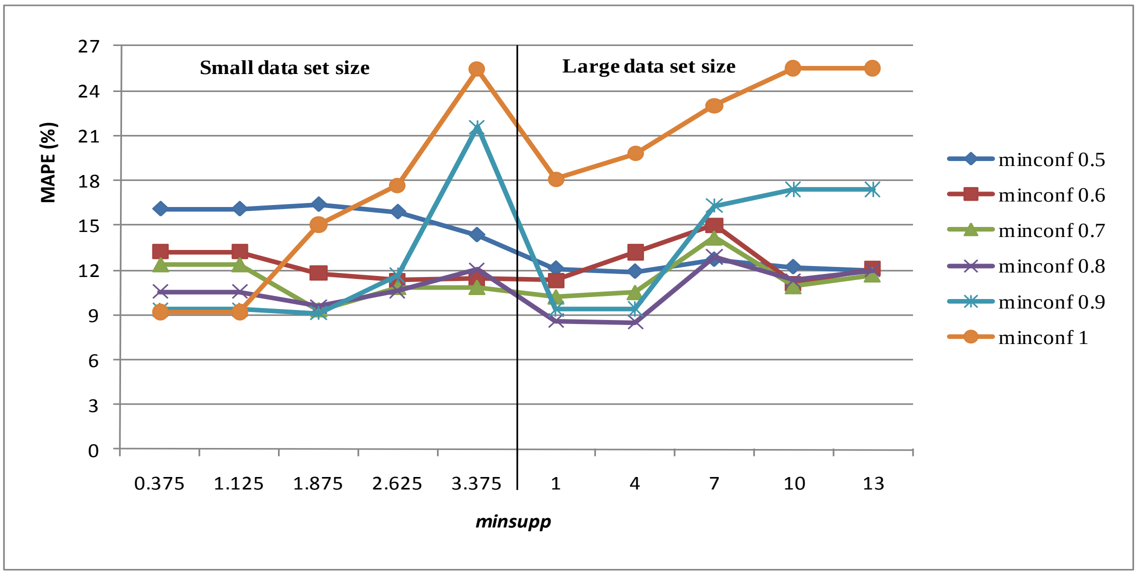


Figure 3.11 The MAPE over different *minsupp* and *minconf* of both small and large data set.

In order to measure the error and evaluate the prediction validity, the predicted values can be compared to the real values. Table 3-13 shows the difference between two case studies with different sizes. It is noted that the MAPE values for the road traffic data set are smaller than for the large size. Also, other criteria measures as mentioned in Section 2.16 are applied in this study for prediction evaluation.

Table 3-13 Statistical gauges of the prediction evaluation

Statistical gauges	Values of a small data set size	Values of a large data set size
MAPE (%)	9.11%	8.5%
NRMSE (%)	11%	9%
NMAE (%)	8%	7%
Uncorr	0.06	0.04

From Table 3-13 the NMAE is very low and the Uncorr value is very close to zero, this implies that the quality of the prediction is reasonably strong.

3.5 The Proposed Second Prediction Model

3.5.1 The Model Description

The use of single *minsupp* for a whole database assumes that all items in the database have the same frequency. However, in real applications, the database contains some items of a high frequency, while others are of a low frequency. The human expert, based on domain knowledge, can set *minsupp* for a specific value in order to find the frequent itemsets. In that case, if *minsupp* is set too high it will extract a low number of frequent itemsets. Thus, the rare items problem will appear and cause a dilemma (called rare item problem). On the other hand, if *minsupp* is set too low, it will extract a high number of frequent itemsets, which causes combinatorial explosions, i.e. all the possible associations will be found. Hence, some of the frequent itemsets are uninteresting or insignificant (Liu et al., 1999, Kiran and Reddy, 2009).

The approaches mentioned in Section 3.2 suffer from one or more of the following problems:

- The single *minsupp* approach is not fair when using a single *minsupp* for the whole data. Single *minsupp* assumes the same frequency for all items in the data. In this manner, the real application data possesses some items with a high frequency; while others possess a low frequency. Therefore, by using a single *minsupp* some of the significant association rules could be missed.
- The use of a threshold measure such as *minconf* to assess the rules is not effective. It depends on selecting and tuning to a parameter threshold to extract the best rules.
- Association rules mining techniques produce many association rules that will affect the prediction results.
- Extraction of representative rules may assist in understanding the perspective points of the user, but, it is not dependable and coverable in case of prediction accuracy.
- The use of a human expert to identify the fuzzy sets is not an effective method.

To overcome the dilemma of the rare item problem, Liu et al. (Liu et al., 1999) proposed an algorithm called MSapriroi based on multiple minimum support thresholds approach using Minimum Item Support (MIS), where the number of generated rules depends on the control parameters used.

It is believed that such an approach (i.e. multiple minimum support thresholds) is applied for extracting FARs, then FARs are utilized for a future value prediction. Actually, the diversity within the FARs are developed and considered for their importance in building a reliable prediction model. The prediction model can therefore accept any input data for prediction of a future value. The best FARs (FARs that satisfied minimum threshold value), in some cases, do not always lead to good results. Therefore, the diverse FARs may offer better results.

The proposed second prediction model utilises this concept building a Diverse Fuzzy Rule Base (DFRB). This model considers the trade-off between prediction error and FARs diversity to provide robustness in fuzzy rule base. This approach is based on selecting the strong FARs in order to minimize prediction error and the FARs diversity to maintain the representative rules. The proposed second prediction model utilizes a FCM and adapts a multiple minimum support approach (MSapriori) (Liu et al., 1999) for extracting FARs of rare and highly frequent termsets from fuzzy data set. It is worth mentioning that the MSapriori algorithm is adapted in order to deal with fuzzy data to generate FARs. Figure 3.12 illustrates the proposed second prediction model of DFRB approach which consists of four phases:

First phase: Generating the FARs based on the FCM and MSapriori approach. FCM is applied as an automatic system which transforms the quantitative data set into fuzzy sets (terms). The MSapriori approach is employed to extract the FARs by setting the control parameters, used in MSapriori, suitable values.

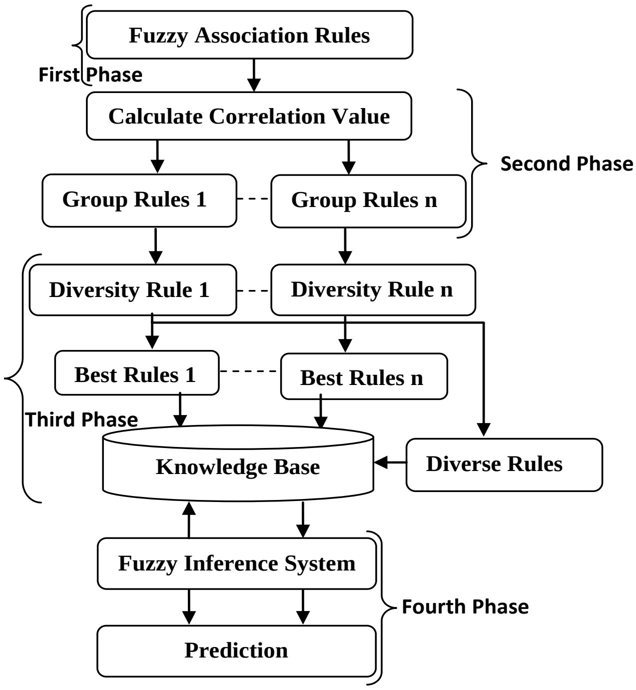


Figure 3.12 The proposed second prediction model with DFRB approach phases.

Second phase: Calculating the correlation coefficient value for each FAR. Correlation values are calculated for each of the FARs using Equations 3-3 and 3-4. The correlation measure is used to evaluate the importance and strength of the association rule, which has been used in (Pach et al., 2008) in order to filter a large number of association rules.

$$\mathrm{corr}(\mathrm{X.FS} \rightarrow \mathrm{Y.FS}) = \frac{\mathrm{supp}(\mathrm{X.FS} \cap \mathrm{Y.FS}) - \mathrm{supp}(\mathrm{X.FS}).\mathrm{Supp}(\mathrm{Y.FS})}{\sqrt{\mathrm{supp}(\mathrm{X.FS}).\left(1 - \mathrm{supp}(\mathrm{X.FS})\right).\mathrm{supp}(\mathrm{Y.FS}).(1 - \mathrm{supp}(\mathrm{Y.FS}))}}$$

(3-3)

where, $\mathrm{corr}(\mathrm{X.FS} \rightarrow \mathrm{Y.FS})$: correlation value of the FAR $(\mathrm{X.FS} \rightarrow \mathrm{Y.FS})$, the interval of its range is [-1, 1]. $\mathrm{supp}(\mathrm{X.FS} \cap \mathrm{Y.FS})$: support value of the FAR, whereas the FARs are formed from the frequent itemset $\{\mathrm{X.FS}, \mathrm{Y.FS}\}$. $\mathrm{supp}(\mathrm{X.FS})$: support value of the antecedent ("IF") part of the FAR. $\mathrm{supp}(\mathrm{Y.FS})$: support value of the consequent ("Then") part of the FAR.

$$\mathrm{Supp}(\mathrm{X.FS}) = \frac{\sum_{i=1}^{N} \prod \mathrm{X.FS_i}}{\mathrm{N}}$$

(3-4)

where, $\mathrm{supp}(\mathrm{X.FS})$:support value of a frequent itemset $\{\mathrm{X.FS}\}$.

The following example explains the calculation of the correlation value. Let $\mathrm{r1: X.Low} \rightarrow \mathrm{Y.Medium}$ be a FAR and its data set is shown in Table 3-14.

Table 3-14 Fuzzy data set.

X.Low	Y.Medium
0.3	0.7
0.9	0.5
0.8	0.2
0.7	0.9

$$\mathrm{corr(X.Low \rightarrow Y.Medium)} = \frac{0.3625 - (0.5075).(0.3975)}{\sqrt{(0.5075).(0.4925).(0.3975).(0.6025)}} = 0.6571$$

$$\mathrm{Supp}(X.FS) = \frac{(0.3).(0.7) + (0.9).(0.5) + (0.8).(0.2) + (0.7).(0.9)}{4} = 0.3625$$

The FARs with positive correlation values are considered and grouped according to their length size $\mathrm{K_i}$. The FAR length is determined by the number of attributes included. For example,

group1 $\mathrm{r1: X.Low \rightarrow Y.Low}$
$\mathrm{r2: Z.Low \rightarrow Y.Medium}$

group2 $\mathrm{r1: V.Medium\ and\ X.Low \rightarrow Y.Medium}$
$\mathrm{r2: W.Medium\ and\ Z.Medium \rightarrow Y.High}$

The correlation value is normalized for each FAR within the group as follows:

1. Determine the maximum correlation value $\mathrm{MaxCorr_K}$
2. Divide each FAR correlation value by $\mathrm{MaxCorr_K}$

For example, $\mathrm{MaxCorr_K} = 0.3$ and two FARs in the group

$$MaxCorr_K(r_1) = 0.3 = 1$$
$$MaxCorr_K(r_2) = 0.2 = 0.66$$

This normalization is applied to rescale the correlation value for each FAR in each group. The reason behind the length size is that the longer the FAR size increases, the smaller the correlation value. As a result, this assists in selecting the best FARs from each group that satisfies the minimum correlation threshold *minCorr*.

The motivations behind using the correlation measure are:

- There are some identified difficulties in tuning and setting the *minsupp* and *minconf* (Pach et al., 2008).
- The correlation measure is used as a filtration for the generated FARs after using MIS *minconf*.
- The correlation measure is robust compared with the confidence measure of Equation 3-5 (as mentioned in (Khan et al., 2008)).

$$\mathrm{Conf}(\mathrm{X.FS}, \mathrm{Y.FS}) = \frac{\mathrm{Supp}(\mathrm{X.FS}, \mathrm{Y.FS})}{\mathrm{Supp}(\mathrm{X.FS})} \tag{3-5}$$

Third phase: Finding the diversity of FARs by calculating the distance between the FARs. A diversity of FARs is calculated for each of the FARs groups, it can be found through clustering FARs within each group as follows:

1. The FARs are clustered based on their distance using Equations 3-8 and 3-9.
2. Clustering the FARs based on their distance, hence the number of FARs and their similarities can be identified within a cluster and other clusters using Equations 3-6 and 3-7; these equations were applied (Deb, 2001) to maintain diversity within the sharing function technique in multi-objective optimization.

$$\mathrm{Range}_{\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{Accum}(\mathrm{Distance}_{\mathrm{ij}}) \tag{3-6}$$

where, $\mathrm{Range}_{\mathrm{i}}$: the value that can find out whether the cluster contains only one FAR ($\mathrm{Range}_{\mathrm{i}}$ enumerates the number of similar rules (close similar) to a rules i); if the $\mathrm{Range}_{\mathrm{i}}$ value equal to 1 or the cluster contains more than one FAR (crowded cluster) if the $\mathrm{Range}_{\mathrm{i}}$ value greater than 1. $\mathrm{Accum}(\mathrm{Distance}_{\mathrm{ij}})$: the accumulative distance between i and all other FARs in the same group, whereas $\mathrm{i}, \mathrm{j} \in$ FARs.

$$\mathrm{Accum}\left(\mathrm{Distance_{ij}}\right) = \begin{cases} 1 - \left(\dfrac{\mathrm{Distance}}{\sigma}\right), \mathrm{if\ Distance} \leq \sigma; \\ 0, \mathrm{otherwise.} \end{cases}$$

(3-7)

where, σ: the threshold value that represents the cluster size (cluster radius), its range is in the interval [0.1, 1]. $\mathrm{Distance_{ij}}$: the distance between two individual FARs i and j, whereas $\mathrm{i, j} \in$ FARs. Measuring the distance between two FARs is the main part at this phase using Equations 3-8 and 3-9, these FARs can be clustered based on their similarities. The similarity between two FARs can be found as follows:

$$\mathrm{S_{ij}} = \frac{\left|\mathrm{RAFS_i} \cap \mathrm{RAFS_j}\right|}{\left|\mathrm{RAFS_i} \cup \mathrm{RAFS_j}\right|}$$

(3-8)

where, $\mathrm{S_{ij}}$: the similarity between two FARs (i and j), the interval of its range is [0, 1]. $\mathrm{RAFS_i}$ (Rule Attribute Fuzzy Set): the fuzzy set concerning an attribute within the $\mathrm{FAR_i}$. $\mathrm{RAFS_j}$: the fuzzy set concerning an attribute within the $\mathrm{FAR_j}$. Consequently, the distance can be calculated using Equation 3-9.

$$\mathrm{Distance_{ij}} = 1 - \mathrm{S_{ij}}$$

(3-9)

The following example illustrates the calculation of the distance. Assuming that $\mathrm{FAR_i}$ and $\mathrm{FAR_j}$ are two rules as follows:

$$\mathrm{FAR_i}: X.Low\ and\ Z.Medium \rightarrow Y.Low$$
$$\mathrm{FAR_j}: X.Medium\ and\ Z.Medium \rightarrow Y.Low$$

Then, the similarity between $\mathrm{FAR_i}$ and $\mathrm{FAR_j}$ can be calculated by using Equation 3-8:

$$\mathrm{S_{ij}} = \frac{|2|}{|4|} = 0.5$$

The numerator 2 comes from $\mathrm{FAR_i.[Z.Medium]}$ similar to $\mathrm{FAR_j.[Z.Medium]}$ and $\mathrm{FAR_i.[Y.Low]}$ similar to $\mathrm{FAR_j.[Y.Low]}$, whereas denominator 4 comes from all attributes fuzzy sets $\mathrm{FAR_{i,j}.[X.Low, X.Medium, Z.Medium, Y.Low]}$.

Once the similarity between FAR_i and FAR_j is calculated, then the distance can be found by using Equation 3-9, below:

$$\text{Distance}_{\text{ij}} = 1 - 0.5 = 0.5$$

Fourth phase: Selecting the best FARs with the highest correlation values and the diverse FARs, then, storing these FARs in the KB to be used by the FIS for a prediction. The approach for extracting robust and diverse fuzzy rule base is summarized in Figure 3.13:

Input: Fuzzy Association Rules (FARs), minimum Correlation threshold value *minCorr*, σ value, number of the Diverse Rules (DR).

Output: Robust and Diverse Fuzzy Rule Base.

Method:
1. Calculate the correlation value corr for each FAR_i
2. Divide FAR into a group based on their length size K_i
3. Sort (rank) FARs automatically in each group K_i based on their highest correlation value.
4. Select the best FARs from each group K_i
   For each K_i
       IF $corr(\ FAR_i) \geq minCorr$
           $Best_{FAR_i} \cup FAR_i$
       EndIF
   EndFor
5. Identify the diverse FARs for each group K_i
   $N\_Range_i\_1$=*counting the number of* $FARs$ *of* $Range_i(Range_i\ eqaul\ 1)$.
   $DR \leq N\_Range_i\_1$.
   For each K_i
       IF $(corr(\ FAR_i) < minCorr)$
           While $(\left|Diverse_{FAR_i}\right| \geq DR)$
                   IF $(Range_i(FAR_i) == 1)$
                       $Diverse_{FAR_i} \cup FAR_i$
                   EndIF
           EndWhile
       EndIF
   EndFor

Figure 3.13 Extraction of robust and diverse fuzzy rule base algorithm.

Figure 3.14 shows an example of the selection of the fuzzy rule base with robustness and diversity.

Group1 of FARs, length size $K_i = 2$	$corr$ value	$Range_i$
$A.Low \rightarrow Y.\mathrm{Low}$	1	1
$B.Medium \rightarrow Y.Low$	0.9	1
$B.High \rightarrow Y.High$	0.8	1
$C.Low \rightarrow Y.Medium$	0.6	1
$D.High \rightarrow Y.Low$	0.5	1

Group 2 of FARs, length size $K_i = 3$	$corr$ value	$Range_i$
$A.Low\ and\ \ B.Low \rightarrow Y.Medium$	0.9	1
$A.High\ and\ B.Medium \rightarrow Y.High$	0.8	1.7
$B.Low\ and\ \ C.Medium \rightarrow Y.Low$	0.6	2
$B.Medium\ and\ C.High \rightarrow Y.High$	0.5	1
$B.High\ and\ \ D.High \rightarrow Y.Medium$	0.4	1.8
$C.Low\ and\ \ D.High \rightarrow C.Medium$	0.4	1

Assuming that *minCorr*=0.7, and DR=1

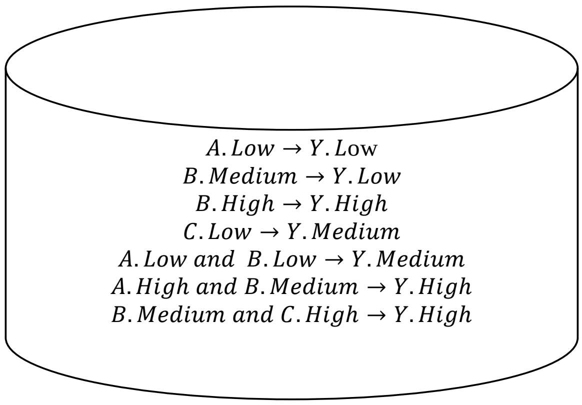


Figure 3.14 Selection of the fuzzy rule base with robustness and diversity.

3.5.2 Experimental Results and Comparative Study

In order to demonstrate the technical feasibility of the proposed KD model (second prediction model of DFRB approach), the model discussed in the previous sub-section is applied in two data sets. The first data set is applied in a quantitative data set called Abalone data set and the second is applied in a quantitative road traffic data set (100 records).

The proposed model has been applied to predict the ATT in road traffic data and the Abalone ring that represents the Abalone age in Abalone data.

Further analysis of the Abalone data to detect an outlier data was undertaken as follows:

- Calculating mean value and standard deviation for each attribute of the data set called $dataset_n$, where n is the data set size.

- Constructing two data matrices. First matrix represents a repeated mean value in each record called $mean_n$; its size equals the data set size. The second matrix represents a repeated standard deviation value in each record called std_n; its size equals the data set size.
- Finding outlier data by using the following equation:

$$outlier = |dataset_n - mean_n| > (3 * std_n) \tag{3-10}$$

The Equation 3-10 considers the data value as an outlier data, when the absolute value that represents the difference between each attribute value of the data set and its mean value is to be 3 times greater than its standard deviation value (3 is an assuming value, which is a threshold and can be changed). The role of using standard deviation is to evaluate the data set distribution (dispersion) from its mean value. The more distribution (spread or outlier) in data, the higher the deviation. The outlier data are kept to ensure that the proposed KD model is working in case of noisy data (unbalanced data distribution).

The same analysis that is applied to the Abalone data is also adapted in the road traffic data for detecting an outlier data, but the noisy data found in road traffic data is less than that which exists in Abalone data. Therefore, these outlier data are kept to ensure the proposed KD model is working in case of noisy data (unbalanced data distribution).

For the purpose of evaluation and validation, prediction quality is assessed using one of the statistical measures called Mean Absolute Percentage Error (MAPE), highlighted in Equation 2-8. Table 3-15 shows the sensitivity of MAPE adapting *minCorr* over different numbers of FARs and Table 3-16 shows the sensitivity of MAPE adapting *minCorr* over different numbers of FARs and the diverse FARs, where $\sigma = 0.3$ in both tables. This value is selected based on different experiments that have been conducted. The results shown in Table 3-15 and Table 3-16 clearly demonstrate that MAPE is not affected when the diverse FARs are added.

Table 3-15 The sensitivity of MAPE (%) adapting *minCorr*.

minCorr	Number of FARs	MAPE (%)
0.4	34	22.2%
0.5	22	9.3%
0.6	**17**	**8.1%**
0.7	15	10.6%
0.8	12	14.2%
0.9	8	23.6%

Table 3-16 The sensitivity of MAPE (%) adapting *minCorr* and diverse FARs.

minCorr	Number of FARs	MAPE (%)
0.4	34	22.2%
0.5	23	9.3%
0.6	**18**	**8.1%**
0.7	17	10.6%
0.8	14	14.2%
0.9	10	23.6%

Table 3-17 represents the sensitivity of MAPE adapting σ value (cluster size) over different numbers of diverse FARs when considering 17 FARs. It can be easily noted that when the value of σ is less than 0.3 or greater than 0.6, it does not affect MAPE, however, when it is between 0.3 and 0.6, MAPE changes slightly. In addition, when the value of σ is between 0.3 and 0.6 then several numbers of diverse FARs are generated which slightly affect the MAPE of the testing data. However, when the selection of σ is less than 0.3, each cluster may include one FAR; and when the value of σ is greater than 0.6, the cluster may include all FARs.

Table 3-17 The sensitivity of MAPE (%) and σ value of *minCorr*= 0.6 with diverse FARs.

σ vlaue	Number of diverse FARs	MAPE (%)
0.1	-	8.1%
0.2	-	8.1%
0.3	**1**	**8.1%**
0.4	**3**	**8.1%**
0.5	4	9.9%
0.6	2	9.9%
0.7	-	8.1%
0.8	-	8.1%
0.9	-	8.1%
1	-	8.1%

The aim of the inclusion of diverse (representative) rules is to cover low frequency data. In addition, the experimental results intend to ensure that diverse rules do not significantly increase (effect) the prediction error. It is observed from Table 3-15, Table 3-16 and Table 3-17 that, the inclusion of a suitable number of diverse rules does not have a significant effect on the prediction error. Therefore, the diversification method is integrated with the proposed hybrid prediction FACRM model in the next chapter (more detail in Chapter 4).

Figure 3.15 shows the sensitivity of the MAPE and *minCorr* for the brute-force approach (the diversity is employed in all FARs together of different K_i (length size of FAR)). The graph shows that when *minCorr* is 0.49, the minimum MAPE is 11.2%, and it contains rules that cover most cases. It is noted that as *minCorr* decreases from 0.49, the MAPE increases. This is explained by producing a large number of rules (decrease in *minCorr* is accompanied by an increase in the uninteresting FARs to cause noise for the FIS). Moreover, it is observed that when the correlation value increases beyond 0.49, the MAPE also increases. Again this is explained by producing a small number of FARs, which do not give robust results for the FIS (the increase in *minCorr* implies a decrease in the number of relevant rules).

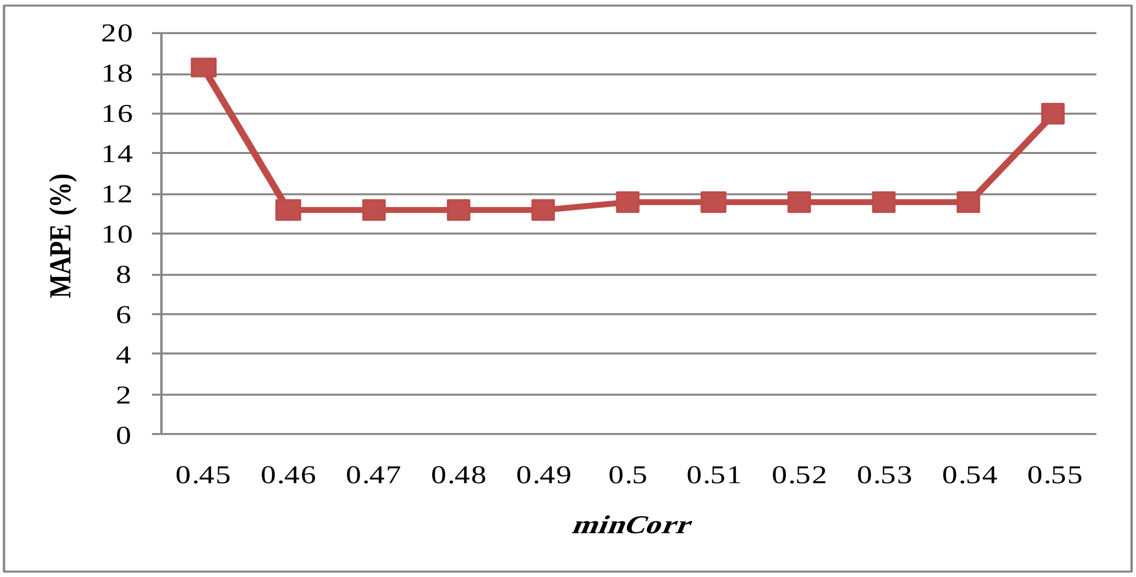


Figure 3.15 The sensitivity of MAPE and *minCorr*.

Table 3-18 depicts the sensitivity of MAPE over different numbers of FARs and diverse FARs, where $\sigma = 0.3$ and *minCorr* = 0.49, which illustrates the brute-force approach. This approach produces a good result, its MAPE is acceptable compared with the employed diversity in FARs with different individual K_i (length size of FAR), as shown in Table 3-18. The diversity is applied in FARs with different K_i, they produce 8.1% MAPE and generate a slightly high number of FARs when compared with the brute-force approach, which produces 11.2% MAPE and less numbers of FARs.

Table 3-18 The sensitivity of MAPE (%) over different numbers of FARs and diverse FARs.

Number of FARs	Number of diverse FARs	MAPE (%)
11	0	11.2%
11	1	11.2%
11	2	11.2%
11	3	18.3%

Other measures as mentioned in Section 2.16 are applied in this study to evaluate the predicted quality. The values of such criteria are shown in Table 3-19.

Table 3-19 Calculation of the evaluation criteria.

Measure criteria	Road traffic data	Abalone data
MAPE (%)	8.1%	25.5%
NRMSE (%)	9.5%,	37.2%
NMAE (%)	6.7%,	27.8%
Uncorr	0.05	0.51

Table 3-19 shows the results using different criteria measurements. It is noted that different criteria yield different values. This is attributed to the different ranges of values in each data set.

The results are compared to those of the integrated FCM and Apriori approach. Prediction quality is assessed using MAE of Equation 2-10 and MAPE of Equation 2-8. The experimental results on the Abalone and road traffic data are summarized in Table 3-20.

Table 3-20 Calculation of MAE and MAPE (%).

Data set	FCM and Apriori	FCM and DFRB
Abalone	3.77 (Lu et al., 2003a)	2.78
Road traffic	9.1% (Section 3.4.2.2)	8.1%

Table 3-20 presents the results from using two different methodologies, the first concerns the integrated FCM and Apriori, while the second deals with FCM and DFRB. The experimental results show an improvement of future value predictions by minimizing MAE and MAPE using the proposed second model that employed the DFRB approach. The results are compared with reported work in (Lu et al., 2003a) and (Section 3.4.2.2), and the proposed second model of DFRB approach gives better results than the results yielded from (Lu et al., 2003a) and the first model (Section 3.4.2.2). It is noted that the result in (Lu et al., 2003a) was generated by two methods. The first method applied FCM and Apriori algorithms, and then Genetic Algorithm (GA) was applied for tuning the fuzzy sets where its MAE was 3.77, while the value of MAE using the proposed model of DFRB is equal to 2.78. The second method presented in (Lu et al., 2003a) proceeded as the first method, but it used variable thresholds in the prediction in order to minimize the MAE. The MAE, based on this method, is equal to 1.77.

The proposed model of DFRB produces a good result and can be used with a wider data set (even when the data set contains noisy data). This provides more generalized prediction method.

3.6 Summary and Conclusion

This chapter has presented two prediction models. The first model is based on the single support value threshold, which has been tested for two (small and large) data sets in the road traffic domain taken from the road traffic simulation model. It is noted from the results that the model has effectively minimized MAPE, which is sensitive to *minsup* and *minconf* values. It is also noted that a large data set size offered lower MAPE compared to the small data set.

The model used Fuzzy C-Means (FCM) to determine centres for each field independently from the whole field. It is noted that, if the whole data is used, FCM may cause an overlapping problem to fuzzy sets (membership functions). However, this problem can be solved by using an optimization technique. In addition, Apriori algorithm used a single *minsupp* for the whole database; for instance to consider and assume the same frequency for all items in the data set. In this manner, the real application data possessed some items of a high frequency, while other items possessed low frequency. To overcome this issue, multiple support approach and diversification method were considered for a proper solution as in the second model.

For the second model, an approach was proposed entitled Diverse Fuzzy Rule Base which focused on the Fuzzy Association Rules (FARs) extraction and effective selection to predict the future values. The approach includes both the significant and diverse FARs through its capability in filtering them in order to extract the best FARs, which impacts the prediction quality. This approach generates the FARs, and then calculates the correlation coefficient value for each FAR. After that, it clusters the FARs to determine its diversity while selecting the best FARs (with a higher correlation value) and diverse FARs for prediction using Fuzzy Inference System (FIS). The proposed approach expresses and maintains the trade-off between both the prediction error and the FARs diversity. The approach is employed to support the second model. The model was applied to two case studies: a data set related to a road traffic domain, and the Abalone data set. It is noted that the second model of the DFRB approach offers less prediction error as compared to a technique reported in the literature.

In the next chapter, a proposed hybrid model based on an improved multiple support associative classification approach will be described in detail to explore further levels of performance and minimum prediction error.

CHAPTER FOUR

4 PREDICTION MODEL WITH IMPROVED MULTIPLE SUPPORT ASSOCIATIVE CLASSIFICATION APPROACH

4.1 Chapter Overview

In this chapter, a Fuzzy Associative Classification Rule Mining (FACRM) model is proposed. This model is based on improved multiple supports and Associative Classification (AC) approaches to improve the reliability of prediction by minimizing prediction error. The proposed FACRM is based on four main stages. In the first stage knowledge is discovered through the integration of the improved Gustafson-Kessel (G-K) algorithm and the proposed Fuzzy Associative Classification Rules (FACR) algorithm. The improved G-K algorithm is used as a pre-processing step to transform quantitative data into fuzzy data, while FACR improves current associative classification approaches by adapting the improved multiple support algorithm. The improvement enables the discovery of significant rules, facilitates a direct pruning of unnecessary rules and deals with unbalanced data. Additionally, FACR utilizes a vertical scanning format for a database to improve the performance of the rules extraction process, instead of using multi-level scan database as in Apriori approach. The second stage filters the extracted rules by calculating a correlation value for each rule to select strong rules. In the third stage, the rules are clustered based on the diversification method (see Chapter 3), which can be achieved by measuring the distance between rules. The diversification method is able to extract the significant (best rules) and representative rules which are stored in Knowledge Base (KB). The aim of the representative rules is to cover infrequent data. In the final stage, the KB is utilized in an application domain for prediction.

The proposed FACRM model can provide a generalized prediction model to deal with different application domains. The validation of the FACRM model is conducted using two sets of experiments. The first validation uses different benchmark data sets from the University of California, Irvine (UCI) of machine learning and KEEL (Knowledge Extraction based on Evolutionary Learning) repositories, then the results of FACRM are compared with common prediction models (Artificial Neural Network (ANN), Support Vector Machine (SVM), Stepwise Regression (SR), Classification and Regression Tree (CART)). The experimental results of the first validation show that the proposed model discovers rules that effectively minimize the prediction error rate. Also, the performance of the proposed model is comparable with frequently used and well-known prediction models. In the second validation experiments apply the proposed model to benchmarking of gas furnace data for the Box and Jenkins problem that is widely used for modelling and identification. The experimental results of the second validation confirm that the FACRM model produces comparable and satisfactory results with respect to the results reported in the literature.

4.2 Introduction

Data mining technology contains many tasks/techniques such as classification and association rules. These techniques extract hidden data (knowledge) from a large database for prediction and decision making purposes. Therefore, discovering significant knowledge from a large database remains one of the most important data mining tasks, which is capable of building an accurate prediction model. Much effort and attention is required in this research field in order to build an effective and accurate prediction model.

Classification is a task of categorizing a training data into predefined groups or classes with the aim of building a classifier model. It is also about grouping records in a training data, where each record contains a collection of attributes and one of the attributes is considered as the output (class label) (Thabtah et al., 2005, Pach et al., 2008). The prediction is a task of forecasting a new data (the new input data of an unknown output) based on a current classifier model (learned model).

It can be noticed that most of the current association rules mining approaches have been extracted association rules (knowledge) based on a high frequency occurring. Conversely, the nature of the real-life applications and their data sets are generally inconsistent and have both rare and frequent items. However, the rare items are difficult to be identified and perceived because of their low quantity data. As mentioned in chapter 3, Liu et al. (Liu et al., 1999) proposed an algorithm called MSapriroi to solve the dilemma of the rare item problem.

Several studies have developed specific prediction models for specific domains (Polydoras et al., 1998, Shin et al., 2005, Grivas and Chaloulakou, 2006, Ivanciuc, 2007, Dahou et al., 2009, Cho et al., 2009, Karabatak and Ince, 2009a, Karabatak and Ince, 2009b, Azzeh et al., 2010). However, these models have some practical problems. Firstly, these models are constructed for special domains. Secondly, the models are either statistical models or machine learning models such as Artificial Neural Network (ANN) and Support Vector Machine (SVM), which are classified as non-rule based methods.

The main aim of this research study is to build a generalized prediction model for different application domains. In this chapter, the prediction model developed in the previous chapter is enhanced by considering accuracy and effectiveness, comprehensibility and efficiency. The model is based on the improved multiple support and Associative Classification approaches, entitled Fuzzy Associative Classification Rules Mining (FACRM) model. The accuracy and effectiveness relate to correct and reliable prediction of a future value, while the comprehensibility confirms model's ability to generate intuitive rules that are able to identify domain knowledge, i.e. explicit rules to be clear for a human expert. The efficiency performs an optimization process for scanning a database to extract rules. The proposed FACRM is based on an integration and utilization of different techniques/approaches as follows: fuzzy clustering, the improved multiple support approach, associative classification approach and diversification method. The fuzzy clustering, in particular, the improved Gustafson-Kessel (G-K) clustering algorithm is used for pre-processing a data by transforming it into a fuzzy data. The improved multiple support approach is employed for extracting frequent itemsets (termsets) by using Support Difference (SD), which is effectively able to generate frequent itemsets including rare items and to limit the combinatorial explosion (restrict from insignificant frequent itemsets including frequent items) as mentioned in Section 2.3.4. The associative classification approach is employed for pruning association rules to find Fuzzy Classification Association Rules (FCARs) that contain only an output attribute in a consequent part ("Then" part) of the rules. The diversification method is utilized to find the best and representative rules for further improvement in prediction.

4.3 Related Literature Work

Most of the well-known classification techniques are based on heuristic/greedy strategy approaches such as the decision tree C4.5 (Quinlan, 1993) and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Cohen, 1995). These approaches aim to discover a small sub-set of rules to represent a training data and build a classifier (Thabtah et al., 2006). The greedy approach is employed through a traditional classification technique, which operates in stages. In each stage, the solution taken seems to be efficient, without considering the next stages. However, these techniques which play a vital role in some cases, suffer from comprehensive rules. The rules generated by these techniques are of a different nature as they are hardly understandable by the user (Thabtah et al., 2005, Pach et al., 2008).

A variety of techniques have been developed through the integration of association rules and classification. These are known as Associative Classification (AC) such as Classification Based on Associations (CBA) (Liu et al., 1998), Multi-class Classification based on Association Rules (MCAR) (Thabtah et al., 2005), and Combination Strategy for Multi-class Classification (CSMC) (Liu et al., 2008b). However, the rules that are extracted by AC techniques cannot be discovered through the use of traditional classifications. Therefore, the integration of association rules and classification is a promising field of research. The first approach which adapted association rules for classification purposes was proposed by Liu et al. (Liu et al., 1998) in the form of the so-called CBA, which applied the popular Apriori algorithm. Their approach was able to extract CARs, with each rule belonging to a specific class label (output attribute). CARs should satisfy both *minsupp* and *minconf*. An AC technique called MCAR was developed by Thabtah et al. (Thabtah et al., 2005). This technique used a single *minsupp* and utilized a vertical format for scanning a database and extracting associative classification rules. As a result, MCAR is deemed to be more accurate than CBA for the use of a proper rules ranking method. An associative classification approach was proposed by Pach et al. (Pach et al., 2008). It uses the fuzzy approach for a data discretization.

A classification approach based on association rules was proposed in (Vo and Le, 2009). This approach generates the associative classification rules using a tree structure called Equivalence Class Rule (ECR). ECR constitutes a classifier by extracting associative classification rules and pruning redundant rules. It scans a database one time to calculate support value for the itemsets by making an intersection between identification cases (a row's number).

Intelligent techniques, such as ANN and SVM, are considered to be effective models, and have been applied in different domains for the purpose of predicting future values (Sapankevych and Sankar, 2009, Karabatak and Ince, 2009a). An integration of different Artificial Intelligence (AI) models can achieve better results, by combining the strengths of AI models and reducing the weaknesses associated with using only a single model. Cheng and Roy's (Cheng and Roy, 2010) prediction model called Evolutionary Fuzzy Support Vector Machine Inference Model, for Time Series Data which integrates AI techniques, such as genetic algorithm, fuzzy logic and support vector machine. This model is used to predict the cash flow to help control the management of projects. In this model, Fuzzy Logic was used as a weight for the support vector machine, while the Genetic Algorithm was applied to optimize the parameters of the fuzzy sets and the weights of support vector machine. The model in (Cheng and Roy, 2010) was evaluated by the Root Mean Square Error (RMSE) and the Average Percentage Error (APE). An expert system based on association rules mining and ANN for predicting and detecting breast cancer was presented by Karabatak and Ince (Karabatak and Ince, 2009a). Association rules mining, specifically Apriori algorithm, was applied for feature selection and reducing the dimension of the Wisconsin breast cancer data where the ANN is utilized on the reduced data for predicting scenarios. The evaluation performance of this system (association rules mining and ANN) was compared with

ANN model. Similarly, the same method was applied to detect and predict erythemato-squamous diseases data (Karabatak and Ince, 2009b).

A prediction model of the compressive strength of high-performance concrete (that used in the concrete construction industry) using a back-propagation ANN, was proposed and demonstrated by Yeh (Yeh, 1998). Another model based on a back-propagation neural network was proposed to predict the slump flow of concrete (Yeh, 2007). Evaluation of both these models concluded that using ANN, outperforms the regression analysis on these specific domains.

4.4 The Proposed Fuzzy Associative Classification Rules Mining (FACRM) Model

4.4.1 The Model Description

As discussed above there are many reported researches on the development of prediction models in the literature. However, the reported models still suffer from one or more of the following issues:

- In addition to the issues mentioned in Section 3.5.1. Association rules mining techniques that produce many association rules are difficult to evaluate (this affects the prediction result in an advanced step). Also, the current AC approaches produces some of insignificant rules.
- As discussed in Section 3.5.1 the issue of using the single *minsupp* approach. Single *minsupp* assumes the same frequency for all items in the data. Therefore, it causes the rare items problem. Similarly as mentioned in Section 2.3.4 the problem of setting a low β value (β is a user-defined parameter to control the relation between Minimum Item Support (MIS) value for each item in data and its actual frequency which

can be from 0 to 1), which leads to generating uninteresting frequent itemsets (including frequent items) having low support values. On the other hand, if β value is set to be high, then Mutliple Support Apriori (MSapriori) suffers from generating the frequent itemsets including rare items.

- Apriori fashion works in level-wise search, which needs to scan a database for each iteration (level), thus, the technical performance is reduced.
- As mentioned in Section 3.5.1 the use of *minconf* threshold measure, to assess the rules is not effective.

In this work, an enhancement of the FACRM model for prediction is proposed, which aims to address some issues highlighted above. The enhancement includes:

- The use of a proper method for pre-processing data reflecting on the rules generation process. This can be carried out by using the improved G-K algorithm for pre-processing of a quantitative data.
- The use of an appropriate method to identify Minimum Item Support (MIS), which assigns an actual minimum threshold for each item (attribute) in data and assists in extracting significant rules. This can be achieved by applying and adapting an improved approach called Improved Multiple Support Apriori (IMSapriori), introduced in Section 2.3.4 to find proper MIS for fuzzy data. This can help in building a generalized model which is applied for prediction in several domains.
- The use of an effective method for scanning a database during the rules extraction process by employing Enhanced Fuzzy Data Representation (EFDR) presented in Section 2.10, which helps to improve performance

of rules. It is based on the adaptation of the vertical data representation method to deal with fuzzy data.

- The analysis and selection of significant rules (knowledge) in association rules mining and AC approaches. The significant and diverse rules are selected to find representative rules of a whole data set providing flexibility in building a reliable prediction model. This selection is accomplished based on developing a diversification method to select positive correlation values of rules, and to cluster and calculate distance between rules.

The proposed FACRM model for prediction is shown in Figure 4.1. There are four main phases in the model.

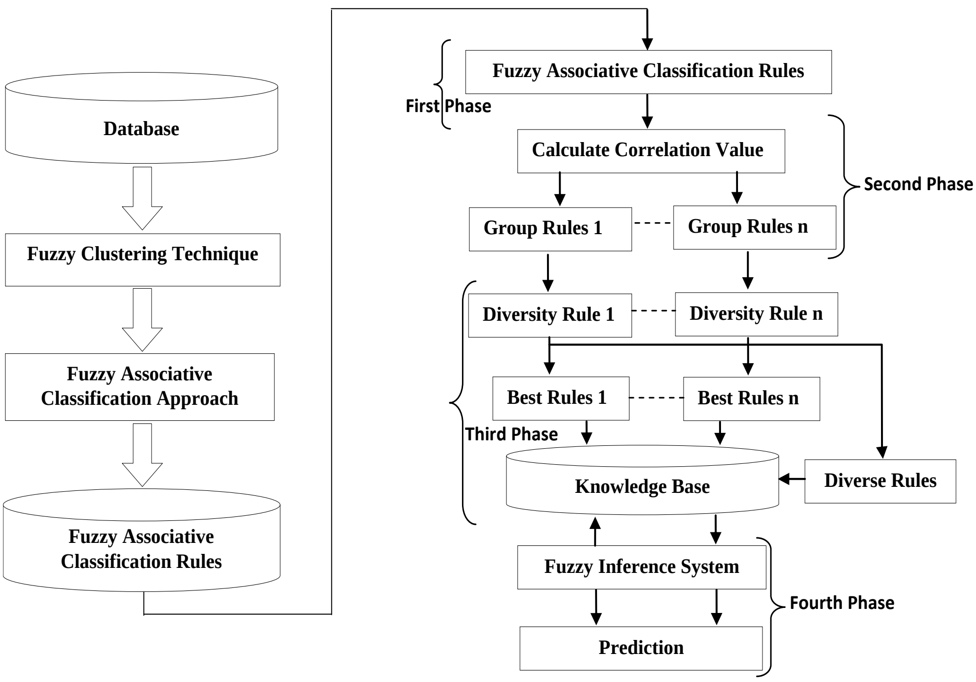


Figure 4.1 The proposed FACRM model for prediction.

First phase: Generating fuzzy classification association rules (or Fuzzy Rule Base (FRB)) from a fuzzy data using the Fuzzy Associative Classification Rules (FACR) algorithm as shown Figure 4.2. The improved G-K algorithm is employed as an automatic system, which transforms quantitative data set into fuzzy data (fuzzy terms).

Definition: Let $\mathrm{FTI} = \{\mathrm{FTi_1}, \mathrm{FTi_l}, \dots, \mathrm{FTi_n}\}$ be fuzzy terms (sets) of input fuzzy data, $\mathrm{FTO} = \{\mathrm{FTo_1}, \mathrm{FTo_j}, \dots, \mathrm{FTo_m}\}$ be fuzzy terms (sets) of output fuzzy data and $\mathrm{FD} = \{\mathrm{U_1}, \mathrm{U_2}, \dots, \mathrm{U_z}\}$ be a set of transactions U in a fuzzy database FD of the fuzzy input terms FTI and the fuzzy output terms FTO, where $1 \le l \le n$ and $1 \le j \le m$. Each transaction U in FD is formed by a set of FTi in FTI where $\mathrm{FTi} \in \mathrm{FTI}$ and FTo in FTO where $\mathrm{FTo} \in \mathrm{FTO}$. A strong Fuzzy Classification Association Rule (FCAR) is defined as $\mathrm{X.FTi} \rightarrow \mathrm{Y.FTo}$ where $\mathrm{X.FTi} \subseteq \mathrm{FTI}, \mathrm{Y.FTo} \subseteq \mathrm{FTO}$ and $\mathrm{X.FTi} \cap \mathrm{Y.FTo} = \emptyset$. i.e. the FCAR as $\mathrm{FTi_1}, \mathrm{FTi_l}, \dots, \mathrm{FTi_n} \rightarrow \mathrm{FTo_j}$, assuming that MFTS is a Minimum Fuzzy Term Support value for a fuzzy term of the input FTi and the output FTo, then the FCAR is accepted if its support value is greater than or equal to the minimum value of MFTS from all fuzzy input and output terms within each FCAR; and its confidence value is greater than or equal to *minconf* threshold:

$\mathrm{support\_value(FCAR)} \ge \mathrm{minimum(MFTS(FTi_1), MFTS(FTi_l) \dots MFTS(FTi_n), MFTS(FTo_j))};$

$\mathrm{confidence\_value(FCAR)} \ge minconf$.

FACR algorithm by employing EFDR is described in Figure 4.2 and demonstrated in the following example:

1. Fuzzy Data FD in Table 4-1 is scanned to calculate MIS, called MFTS, based on Equations 2-5 and 2-6. Each candidate termset (1-termsets) is associated with its FD, support value and MFTS as it is shown in Fuzzy Term Set ($\mathrm{FTS_1}$) of Table 4-2. Any fuzzy term $\mathrm{FT_{il}}$, $\mathrm{FT_{oj}}$, where $\mathrm{FT_{il}} \in \mathrm{FTI}$ and $\mathrm{FT_{oj}} \in \mathrm{FTO}$ that fails to pass MFTS is discarded, i.e. the 1-termsets pass their MFTS are stored in Enhanced Fuzzy Term Set $\mathrm{EFTS_1}$ (the marked in Bold Italic). An association of any input 1-termset $\mathrm{FT_{il}}$ with any output 1-temset ($\mathrm{FT_{oj}}$) in $\mathrm{EFTS_1}$ is checked to form 1-associative termset as shown in Table 4-3 of Fuzzy Rule Term Set $\mathrm{FRTS_1}$. An associative termset will be stored in $\mathrm{FRTS_1}$` if passes its MFTS (the

marked Bold Italic). In other words, each frequent termset in $EFTS_1$ of input fuzzy term (1-termset) FT_{il} is associated with each output termset FT_{oj}, and then the support value is calculated from $EFTS_1$ to form 1-associative termset as in $FRTS_1$, if a termset satisfies its MFTS then it is called frequent associative termset.

Table 4-1 Fuzzy Data (FD).

Case ID	X		Y		Z		O	
	x1	x2	y1	y2	z1	z2	o1	o2
1	0.3	0.7	0.1	0.9	0.9	0.1	1	0
2	0.1	0.9	0	1	0.5	0.5	0.2	0.8
3	0	1	0.3	0.7	0.7	0.3	0.1	0.9
4	0.2	0.8	0.2	0.8	0.75	0.25	1	0
5	0.5	0.5	0	1	1	0	0.5	0.5
6	1	0	0.2	0.8	0.8	0.2	0.4	0.6
7	0.6	0.4	0	1	0.9	0.1	0.3	0.7
8	0.9	0.1	0.3	0.7	1	0	0.75	0.25
9	0.3	0.7	0.5	0.5	0.7	0.3	0	1
10	1	0	0.2	0.8	0.8	0.2	0.5	0.5

Table 4-2 Fuzzy Term Set 1 (FTS_1/*$EFTS_1$*).

LS=0.3, α=0.7								
Case ID	**X**		**Y**		**Z**		**O**	
	x1	***x2***	y1	***y2***	***z1***	z2	***O1***	***O2***
1	***0.09***	***0.49***	0.01	***0.81***	***0.81***	0.01	***1***	***0***
2	***0.01***	***0.81***	1	***0***	***0.09***	0.49	***0.04***	***0.64***
3	***0***	***1***	0.09	***0.49***	***0.49***	0.09	***0.01***	***0.81***
4	***0.04***	***0.64***	0.04	***0.64***	***0.5625***	0.0625	***1***	***0***
5	***0.25***	***0.25***	1	***0***	***0.16***	0.36	***0.25***	***0.25***
6	***1***	***0***	0.04	***0.64***	***0.64***	0.04	***0.16***	***0.36***
7	***0.36***	***0.16***	0.49	***0.09***	***0.36***	0.16	***0.09***	***0.49***
8	***0.81***	***0.01***	0.01	***0.81***	***0.04***	0.64	***0.5625***	***0.0625***
9	***0.09***	***0.49***	0.25	***0.25***	***0.49***	0.09	***0***	***1***
10	***1***	***0***	0.04	***0.64***	***0***	1	***0.25***	***0.25***
Support	***0.365***	***0.385***	0.297	***0.437***	***036425***	0.29425	***0.336***	***0.386***
MFTS	***0.2575***	***0.2775***	0.1895	***0.3295***	***0.2568***	0.1868	***0.2288***	***0.2788***

Table 4-3 Fuzzy Rule Term Set 1 ($FRTS_1$/*$FRTS_1$*`).

1-assocative termset	**Support**
x1o1	***0.2525***
x1o2	0.2375
x2o1	0.2225
x2o2	***0.2875***
y2o1	***0.3255***
y2o2	0.244
z1o1	***0.263***
z1o2	0.272

2. Candidate termsets (2-termsets) are generated from (1-termsets) of EFTS_1 of Table 4-2 then stored in FTS_2, i. e. two termsets are joined of size K from $\mathrm{K}-1$. The support value for each termsets is calculated using Equation 3-4, each termset is associated with its FD as shown in Table 4-4. If the support value of a termset is greater than or equal to its MFTS, then store a frequent 2-termset in EFTS_2. In other words, a training fuzzy data is scanned one time to generate 1-termset involving one term by considering only frequent termsets that satisfy its MFTS of EFTS_1, and then join these frequent termsets from EFTS_1 to form candidate 2-termsets, including two fuzzy terms as in FTS_2 of Table 4-4. Subsequently, keep joining frequent termsets involving more fuzzy terms until EFTS_K is empty.

Table 4-4 Fuzzy Term Set 2 ($\mathbf{FTS_2}$/$\boldsymbol{EFTS_2}$).

Case ID	***x1y2***	x1z1	x2y2	***x2z1***	***y2z1***	***o1***	***o2***
1	***0.27***	0.27	0.63	***0.63***	***0.81***	***1***	***0***
2	***0***	0.03	0	***0.27***	***0***	***0.04***	***0.64***
3	***0***	0	0.7	***0.7***	***0.49***	***0.01***	***0.81***
4	***0.16***	0.15	0.64	***0.6***	***0.6***	***1***	***0***
5	***0***	0.2	0	***0.2***	***0***	***0.25***	***0.25***
6	***0.8***	0.8	0	***0***	***0.64***	***0.16***	***0.36***
7	***0.18***	0.36	0.12	***0.24***	***0.18***	***0.09***	***0.49***
8	***0.81***	0.18	0.09	***0.02***	***0.18***	***0.5625***	***0.0625***
9	***0.15***	0.21	0.35	***0.49***	***0.35***	***0***	***1***
10	***0.8***	0	0	***0***	***0***	***0.25***	***0.25***
Support	***0.317***	0.22	0.253	***0.315***	***0.325***	***0.336***	***0.386***

3. Repeat step 2, until EFTS_K is empty.
4. Candidate termset of 2- associative termset are generated from EFTS_2 by joining each frequent termset with $\mathrm{FT}_{\mathrm{oj}}$ as it is shown in Table 4-5 of FRTS_2, and then prune the candidate associative termsets as follows:

Foreach candidate associative termsets ATS $\in \mathrm{FRTS}_\mathrm{K}$ do
    Foreach $(K-1)$ subset S of ATS do
        IF $(\mathrm{ats}[1] \in \mathrm{S})$or $(MFTS(\mathrm{ats}[2]) = \mathrm{MFTS}(\mathrm{ats}[1]))$Then
            IF $(S \notin \mathrm{FRTS}_{\mathrm{K}-1}{}')\ OR\ (S \notin \mathrm{EFTS}_{\mathrm{K}-1})$Then
                Delete ATS from FRTS_K
            EndIF
        EndIF
    EndFor
EndFor

FACR uses the sorted closure property, i. e. in case a termset is not frequent at $(\mathrm{K}-1)$ termset, it will not be pruned since an increment of terms to it can be frequent. Subsequently, if associative termsets pass their MFTS then store them in $\mathrm{FRTS_2}'$.

5. Repeat step 4, until $\mathrm{EFTS_K}$ is empty.

Table 4-5 Fuzzy Rule Term Set 2 ($\mathrm{FRTS_2}$).

2-associative termset	Support
x1y2o1	0.18115
x1y2o2	0.13585
x2z1o1	0.1541
x2z1o2	0.1609
y2z1o1	0.1904
y2z1o2	0.1346

Frequent termsets and associative termsets are generated recursively from $\mathrm{FTS_K}$ and $\mathrm{EFTS_K}$ respectively, that having a smaller number of termsets $(K-1)$, starting from $\mathrm{FTS_1}$ generated in a single pass of a training fuzzy data. It is worth mentioning that, in case contradiction rules are found then only the rule with the highest confidence value is considered.

The candidate termsets method employed by FACR, scans fuzzy training data to calculate the fuzzy terms supports as in $\mathrm{FTS_1}$. This method is based on an improved multiple support (support difference) of Equations 2-5 and 2-6 and EFDR. In FACR there are no multiple fuzzy data scans. The process of generation of the candidate termsets with their support values using Equation 3-4, is performed only at the previous step. This process reduces the amount of information associated with each step, especially in the case of the availability of high correlated fuzzy data.

Input: Training quantitative data, parameters control: LS, α, *minconf*.

Output: Fuzzy Classification Association Rules (Fuzzy Rule Base (FRB)).

Pre-processing phase:

Transform quantitative attributes into Fuzzy Data (FD) by employing the improved G-K algorithm.

Method:

1. Scan FD to calculate MIS called Minimum Fuzzy Term Support (MFTS) for each Fuzzy Set (Fuzzy Term (FT)) based on Equation 2-5 and 2-6.
2. Find a set of Fuzzy Term Sets (FTS_1) that satisfy their MFTS, then insert the frequent termset into EFTS_1.
3. **Foreach** termset in EFTS_1 Do
4. Join each termset of *fuzzy input term* FT_{il} *and fuzzy output term* FT_{oj}, and store it in FRTS_1
5. **IF** associative termset. Support pass MFTS Then
6. Insert associative termset into FRTS_1`
7. **EndIF**
8. **EndFor**
9. **Foreach** $(\mathrm{K} = 2; \mathrm{EFTS}_{\mathrm{K}-1} \neq \emptyset; \mathrm{K}++)$ **Do**
10. **Foreach** pair of disjoint frequent termsets in $\mathrm{EFTS}_{\mathrm{K}-1}$
11. Join termsets as follows
12. {
13. Insert into FTS_K
14. Select termset; $\mathrm{p.FT}_{\mathrm{i}1}, \mathrm{p.FT}_{\mathrm{il}} \ldots \mathrm{p.FT}_{\mathrm{i(K-1)}}, \mathrm{q.FT}_{\mathrm{i(K-1)}}$
15. From $\mathrm{EFTS}_{\mathrm{K}-1}$ called (p), $\mathrm{EFTS}_{\mathrm{K}-1}$ called (q)
16. Where $\mathrm{p.FT}_{\mathrm{i}1} = \mathrm{q.FT}_{\mathrm{i}1} \ldots \mathrm{p.FT}_{\mathrm{i(K-1)}} \neq \mathrm{q.FT}_{\mathrm{i(K-1)}}$
17. }
18. **EndFor**
19. **Foreach** termset in FTS_K
20. **IF** termset.Support pass MFTS Then
21. Insert into EFTS_K
22. **EndIF**
23. **EndFor**
24. **Foreach** frequent termset in $\mathrm{EFTS}_{\mathrm{k}-1}$
25. Join termsets with $\mathrm{FT}_{\mathrm{oj}}$
26. Insert into FRTS_K
27. **EndFor**
28. **Foreach** associative termset.Support in FRTS_K Do
29. **IF** associative termset. Support pass MFTS Then
30. Insert associative termset into FRTS_K`
31. **EndIF**
32. **EndFor**
33. **EndFor**
34. **Foreach** frequent associative termset in FRTS_K` Do
35. Calculate the confidence value for each frequent associative termset using Equation 3-5.
36. **IF** frequent associative termset $\geq$ *minconf* Then
37. Accept frequent associative termset as FRB
38. **EndIF**
39. **EndFor**

Figure 4.2 FACR algorithm.

Second phase: This phase is the same as the second phase discussed in Chapter 3.

Third phase: This phase is also the same as the third phase discussed in Chapter 3, but measuring the distance between two FRB is different at this phase using Equations 4-1, 4-2 and 4-3 as explained in Section 2.8. These FRB can be clustered based on their similarities. Actually, the similarity between two FRB can be found as follows:

$$\mathrm{S_{ij}} = \frac{\left|\mathrm{RAFS_i} \cap \mathrm{RAFS_j}\right| + \mathrm{FSS_{ij}}}{\left|\mathrm{RAFS_i} \cup \mathrm{RAFS_j}\right|} \tag{4-1}$$

$$\mathrm{FSS_{ij}} = \frac{\left|\mathrm{FS_i} - \mathrm{FS_j}\right|}{\left|\mathrm{MaxFS_{ij}}\right| - 1} \tag{4-2}$$

where, $\mathrm{S_{ij}}$: the similarity between two FRB (i and j), the interval of its range is [0, 1]. $\mathrm{RAFS_i}$ (Rule Attribute Fuzzy Set): the fuzzy set concerning an attribute within the $\mathrm{FRB_i}$. $\mathrm{RAFS_j}$: the fuzzy set concerning an attribute within the $\mathrm{FRB_j}$. $\mathrm{FSS_{ij}}$(Fuzzy Set Similarity): the partial similarity between two fuzzy sets FS (i and j) that belong to the same attribute. $\mathrm{MaxFS_{ij}}$: the maximum number (cardinality) of the fuzzy sets. Consequently, the distance can be calculated using Equation 4-3.

$$\mathrm{Distance_{ij}} = 1 - \mathrm{S_{ij}} \tag{4-3}$$

The following example illustrates the calculation of the distance. Assuming that $\mathrm{FRB_i}$ and $\mathrm{FRB_j}$ are two rules as follows:

$$\mathrm{FRB_i}\text{: X. Low and Z. Medium} \rightarrow \text{Y. Low}$$
$$\mathrm{FRB_j}: X.Medium\ and\ Z.Medium \rightarrow Y.Low$$

Then, the similarity between $\mathrm{FRB_i}$ and $\mathrm{FRB_j}$ can be calculated by using Equation 4-1. Assuming that $\mathrm{MaxFS_{ij}} = 3$ ([1] Low, [2] Medium, [3] High).

$$\mathrm{FSS_{ij}} = \frac{|1-2|}{|3|-1}$$

$$\mathrm{S_{ij}} = \frac{|2| + 0.5}{|4|} = 0.625$$

The numerator $|2| + 0.5$, $|2|$ comes from $\mathrm{FRB_i}$. [Z. Medium] similar to $\mathrm{FRB_j}$. [Z. Medium] and $\mathrm{FRB_i}$. [Y. Low] similar to $\mathrm{FRB_j}$. [Y. Low], 0.5 comes from $\mathrm{FSS_{ij}}$

that represents the partial similarity between two fuzzy sets belong to the same attribute, whereas denominator 4 comes from all attributes of fuzzy sets $\mathrm{FRB_{i,j}}$. [$\mathrm{X.Low, X.Medium, Z.Medium, Y.Low}$].

Once the similarity between $\mathrm{FRB_i}$ and $\mathrm{FRB_j}$ is calculated, the distance can be found by using Equation 4-3, below:

$$\mathrm{Distance_{ij}} = 1 - 0.625 = 0.375$$

Fourth phase: This phase is the same as the fourth phase discussed in Chapter 3.

4.4.2 Experimental Results and Comparative Study

In this sub-section, a data preparation is presented and two validation and comparison methods are conducted. In the first method, the comparison of the proposed model with common prediction models (ANN, SVM, Stepwise Regression (SR) and Classification and Regression Tree (CART)) is demonstrated. In the second method, the identification for the gas furnace time series data is illustrated and compared with the existing literature.

4.4.2.1 Data preparation

The data sets are normalized, which transforms each attribute value in the data set within the range [0, 1]. The attribute values vary across a wide range, therefore, data normalization is used to minimise the effect of these variations, and can also help to avoid numerical complexity through the calculation. The Euclidean norm (unit length) is used for the normalization purpose (Chu et al., 2007, Leopold and Kindermann, 2002) as given in Equation 4-4.

$$\mathrm{X_j} = \frac{\mathrm{X_i}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X_i})^2}} \tag{4-4}$$

where X_j is the normalized value of a data attribute, X_i is an attribute value, n is an attribute size (number of rows).

4.4.2.2 First Validation and Comparison with Common Prediction Models

To validate and assess the effectiveness of the proposed FACRM model, an empirical study is conducted using several data sets as shown in Table 4-6. This represents different domains taken from the University of California, Irvine (UCI) of machine learning (Frank and Asuncion, 2010) and the KEEL (Knowledge Extraction based on Evolutionary Learning) dataset repositories (Alcalá-Fdez et al., 2011). In the test stage, 10-fold cross-validation method is applied. The performance comparison of the proposed FACRM model with frequently used prediction models (ANN, SVM, SR and CART) is shown in Table 4-7, Table 4-8, Table 4-9 and Table 4-10 by using different evaluation criteria. These evaluation criteria are MAPE, MdAPE, RMSE and MAE as explained in Equations 2-8, 2-9, 2-12 and 2-10 respectively.

The sensitivity analysis is carried out to identify the proper values of LS and *minconf* as depicted in Table 4-6.

Table 4-6 Basic information of the data sets from UCI and Keel repositories.

Data set name	Number of features	Number of instances	(Real / Integer / Nominal)	LS		*minconf*
Diabetes	3	43	(3 / 0 / 0)	1	(2.33%)	0.65
Friedman	6	1200	(6 / 0 / 0)	5	(0.42%)	0.45
Plastic	3	1650	(3 / 0 / 0)	1	(0.061%)	0.45
Dee	7	365	(7 / 0 / 0)	15	(4.11%)	0.5
Quake	4	2178	(3 / 1 / 0)	1	(0.046%)	45
Auto MPG6	6	392	(3 / 3 / 0)	15	(3.83%)	0.6
Auto MPG8	8	392	(3 / 5 / 0)	27	(6.89%)	0.6
Weather Izmir	10	1461	(10 / 0 / 0)	155	(10.61%)	0.75
Weather Ankara	10	1609	(10 / 0 / 0)	40	(2.49%)	0.65
Abalone	9	4177	(7 / 2 / 0)	55	(1.32%)	0.6
California Housing	9	20640	(3 / 6 / 0)	100	(0.48%)	0.6
Computer activity	22	8192	(22 / 0 / 0)	2450	(29.91%)	0.7
Treasury	16	1049	(16 / 0 / 0)	100	(9.53%)	0.8
House	17	22784	(17 / 0 / 0)	2000	(8.78%)	0.45

From the results in Table 4-7 and Figure 4.3 , Figure 4.4 and Figure 4.5 it can be observed that FACRM produced less error values than ANN, SVM and SR in terms of MAPE (the mean value of MAPE for all data sets equal to 15.37).

Table 4-7 MAPE (%) values of prediction models for all data sets.

Data set	FACRM	ANN	SVM	SR
Diabetes	**10.63**	13.8	12.12	12.11
Friedman	21.59	**13.17**	22.59	19.80
Plastic	11.1	12.11	15.9	**8.85**
Dee	15.17	14.01	**12.31**	15.53
Quake	**2.31**	7.8	2.45	2.453
Auto MPG6	**11.7**	15.02	14.26	14.93
Auto MPG8	**13.64**	16.11	14.60	16.32
Weather Izmir	**3.52**	7.81	3.85	3.872
Weather Ankara	**5.33**	5.56	9.31	5.38
Abalone	20.61	16.92	**16.87**	19.67
California Housing	**30.94**	43.11	32.36	35.66
Computer activity	**8.84**	131.05	13.28	11.97
Treasury	15.74	**10.11**	18.57	21.91
House	**44.05**	63.81	45.93	64.46
Mean	**15.37**	26.45	16.74	18.06

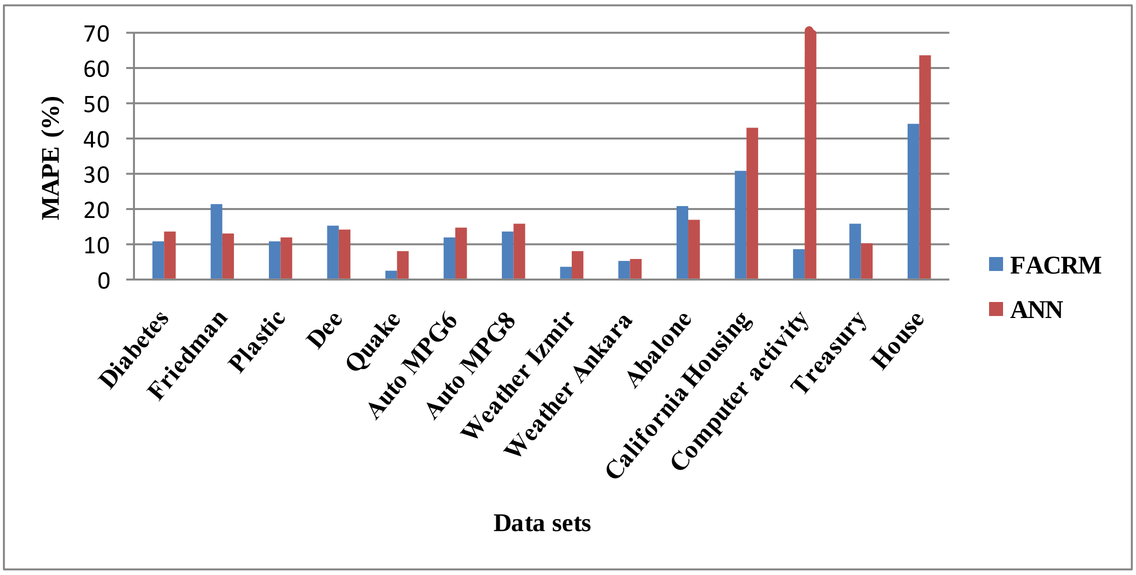


Figure 4.3 The comparison between FACRM and ANN using MAPE values for all data sets.

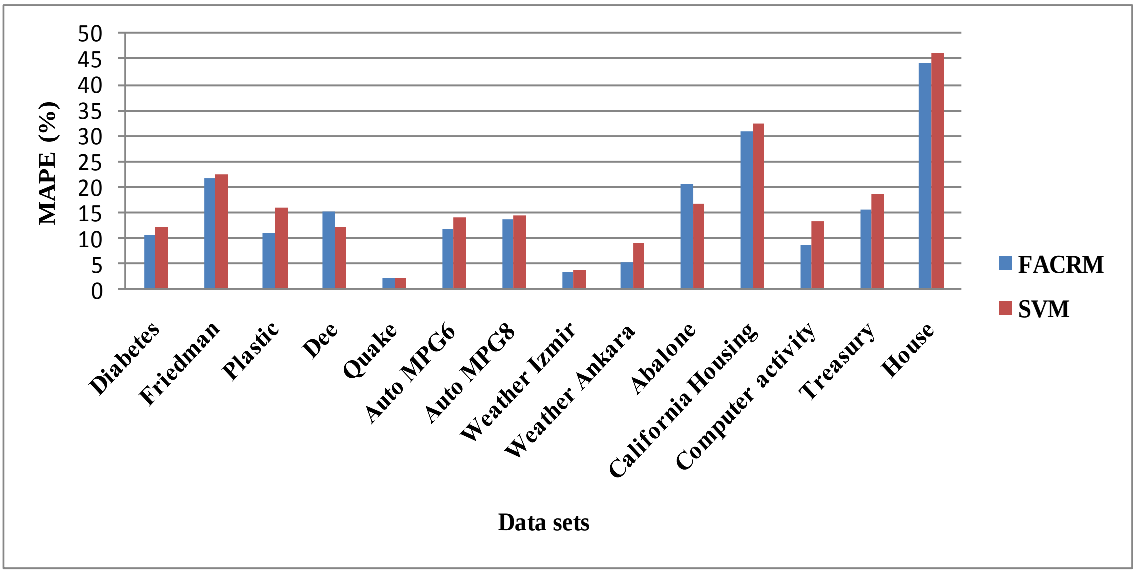


Figure 4.4 The comparison between FACRM and SVM using MAPE values for all data sets.

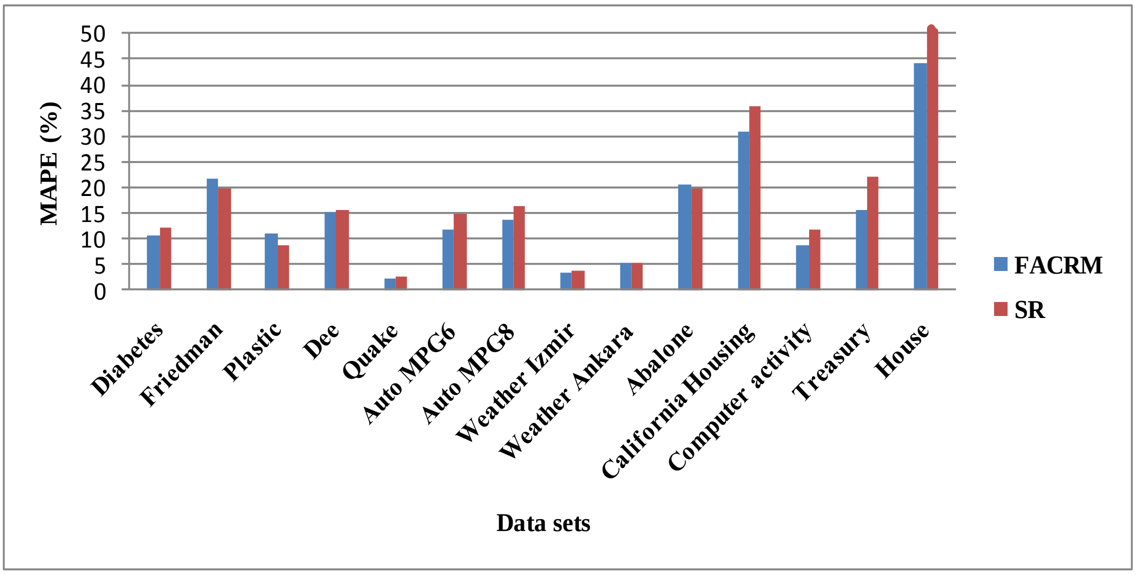


Figure 4.5 The comparison between FACRM and SR using MAPE values for all data sets.

Furthermore, the results in Table 4-8 recommend that FACRM yields to have lower value of errors than ANN, SVM and SR in term of MdAPE (the mean value of MdAPE for all data sets equal to 12.125).

Table 4-9 shows that SR has the minimum value of the mean for all data sets, while SVM has the second minimum value of the mean for all data sets. It is worth mentioning that, FACRM also has a competitive result with SR and SVM using RMSE.

It can be seen from Table 4-9 that there is a slight difference between the mean value of FACRM and SR, which is equal to 0.00015.

Table 4-8 MdAPE (%) values of prediction models for all data sets.

Data set	FACRM	ANN	SVM	SR
Diabetes	**6.56**	11.719	10.07	10.38
Friedman	12.95	**8.74**	14.90	12.53
Plastic	8.49	10.02	13.92	**7.17**
Dee	11.01	**8.76**	10.10	12.86
Quake	**1.59**	7.73	2.14	2.344
Auto MPG6	**9.79**	11.29	12.88	12.98
Auto MPG8	**10.48**	13.18	12.92	12.9
Weather Izmir	**2.99**	6.38	3.41	3.20
Weather Ankara	**3.71**	4.05	8.29	3.90
Abalone	16.6	13.99	**13.4**	16.09
California Housing	25.88	32.66	**24.10**	25.14
Computer activity	**4.10**	14.56	6.48	5.145
Treasury	12.24	**7.84**	14.21	15.65
House	43.36	51.54	**35.7**	45.43
Mean	**12.125**	14.46	13.037	13.265

Table 4-9 RMSE values of prediction models for all data sets.

Data set	FACRM	ANN	SVM	SR
Diabetes	**0.0193**	0.02471	0.01959	0.02041
Friedman	0.00537	**0.003588**	0.005461	0.00504
Plastic	0.003074	0.003224	0.004155	**0.00245**
Dee	0.008326	0.00888	**0.00708**	0.008881
Quake	0.000775	0.001891	0.000718	**0.000677**
Auto MPG6	**0.007361**	0.007812	0.00896	0.007863
Auto MPG8	**0.008476**	0.00881	0.009033	0.00866
Weather Izmir	0.00126	0.00217	**0.001174**	0.001291
Weather Ankara	0.003642	0.003338	0.004122	**0.00264**
Abalone	0.004836	**0.00344**	0.003731	0.00358
California Housing	0.002558	0.00277	**0.002283**	0.00266
Computer activity	0.002581	0.00205	0.0015	**0.00124**
Treasury	0.00717	**0.003131**	0.0075	0.008351
House	0.005256	0.00487	0.004436	**0.004137**
Mean	0.00571	0.00576	0.0057	**0.00556**

From Table 4-10 it can be noticed that the value of the mean of SR also generates the minimum value of MAE equal to 0.00414. However, FACRM also has a comparative result and there is a slightly higher MAE (0.00006) with FACRM with respect to SR. The use of different evaluation criteria such as: MAPE, MdAPE, RMSE and MAE tend to produce different results as obtained in Table 4-7, Table 4-8, Table 4-9 and Table 4-10 which indicates that the error rate of FACRM is satisfactory for the data sets considered.

Table 4-10 MAE values of prediction models for all data.

Data set	FACRM	ANN	SVM	SR
Diabetes	**0.01526**	0.01985	0.01713	0.01708
Friedman	0.00427	**0.00294**	0.0044	0.00390
Plastic	0.0025	0.00268	0.00354	**0.001974**
Dee	0.00669	0.00573	**0.00551**	0.006774
Quake	**0.000508**	0.00166	0.000537	0.000532
Auto MPG6	**0.00558**	0.00649	0.00682	0.00587
Auto MPG8	**0.006421**	0.00723	0.00701	0.00659
Weather Izmir	0.000909	0.0018	0.00095	**0.000781**
Weather Ankara	0.00289	0.002665	0.00459	**0.00204**
Abalone	0.00325	0.00263	**0.00255**	0.00256
California Housing	0.00178	0.00213	**0.00168**	0.00171
Computer activity	0.00131	0.001704	0.001217	**0.000779**
Treasury	0.0048	**0.00263**	0.005527	0.00507
House	0.00266	0.002718	**0.00224**	0.002316
Mean	0.0042	0.00449	0.00455	**0.00414**

Although the experimental results of using different evaluation criteria show that the prediction errors are of different values, the results of overall prediction effectiveness of the proposed FACRM prediction model, is consistently good as shown in Table 4-11. This table summarizes the previous Table 4-7, Table 4-8, Table 4-9 and Table 4-10 of the respective models. Table 4-11 shows the number of the data sets that produced the minimum values for each evaluation criteria. Table 4-11 summarizes all evaluation criteria (MAPE, MdAPE, RMSE and MAE). The results show that the FACRM model performs better for a higher number of data sets than ANN, SVM and SR in terms of MAPE and MdAPE. FACRM has the same number of data sets as that of ANN and SVM using RMSE, and also has the same results as that of SVM and SR in terms of MAE. SR produces lower error for a high number of data sets using RMSE. It can be noticed that, using different evaluation criteria offers diverse results, but overall FACRM has a higher average value among all the data sets using all the evaluation criteria as demonstrated in Table 4-11.

Table 4-11 The data sets numbers that produced minimum error values out of 14 data sets.

Evaluation criteria	FACRM	ANN	SVM	SR
MAPE (%)	9	2	2	1
MdAPE (%)	7	3	3	1
RMSE	3	3	3	5
MAE	4	2	4	4
Average	5.75	2.5	3	2.75

Table 4-12 shows the average of mean error values from Table 4-7, Table 4-8, Table 4-9 and Table 4-10 of all prediction models using all evaluation criteria. It can be seen that, FACRM has the lowest value of the average of mean error among all the prediction models, at 6.88.

Table 4-12 The average of mean error values of all prediction models.

Evaluation criteria	FACRM	ANN	SVM	SR
MAPE (%)	15.37	26.45	16.74	18.06
MdAPE (%)	12.125	14.46	13.037	13.265
RMSE	0.00571	0.00576	0.0057	0.00556
MAE	0.00420	0.00449	0.00455	0.00414
Average	**6.88**	10.23	7.45	7.83
Difference (RMSE-MAE)	0.00151	0.00127	0.00115	0.00142

The importance of using MAE and RMSE is to detect the variation in the error values of prediction. The difference between RMSE and MAE in FACRM equals 0.00151 as illustrated in Table 4-12. FACRM is accurate for most of the cases (individual error values of the prediction), while there are some large error values that affect RMSE. The difference values (difference values between RMSE and MAE, for more details about the relation between RMSE and MAE, see Section 2.16) as shown in Table 4-12 illustrate that FACRM has slightly higher difference value than that of other models. This higher difference between RMSE and MAE is due to some cases of large error values. However, the experiments show that the large error values is not that much since FACRM produced the best results in terms of MdAPE (the minimum mean value of MdAPE for all data sets as shown in Table 4-8). Therefore, FACRM is relatively consistent model. Considering the above results and analysis of FACRM, Table 4-13 shows the number of data sets where FACRM produces better results than the other prediction models using different evaluation criteria.

Table 4-13 The number of data sets out of 14 data sets that FACRM beats other models.

Evaluation criteria	FACRM Vs. ANN	FACRM Vs. SVM	FACRM Vs. SWR
MAPE (%)	10	12	11
MdAPE (%)	10	10	10
RMSE	8	7	7
MAE	9	9	6
Average	9.25	9.5	8.5

So far, the comparison has been with non-rule based models (ANN, SVM and SR). Table 4-14 shows the comparison of FACRM with CART, which shows that the performance of the FACRM is very close to that of CART. The mean value of all evaluation criteria of CART equals 5.742, while the mean value of FACRM equals 6.876, with the difference being 1.13. On the other hand, FACRM has a high average number of data sets that produced less error values for all the evaluation criteria which equals 7.5 out of 14 data sets.

Table 4-14 The comparison of FACRM and CART.

Data set	MAPE (%)		MdAPE (%)		RMSE		MAE	
	FACRM	CART	FACRM	CART	FACRM	CART	FACRM	CART
Diabetes	**10.63**	12.48	**6.56**	11.45	**0.0193**	0.0207	**0.01526**	0.01812
Friedman	21.59	**21**	**12.95**	13.68	**0.00537**	0.00542	**0.00427**	0.004468
Plastic	11.1	**10.6**	**8.49**	9.1	**0.003074**	0.003136	**0.0025**	0.002564
Dee	**15.17**	17.16	**11.01**	11.47	0.008326	**0.006958**	**0.00669**	0.006958
Quake	**2.31**	3.04	**1.59**	2.47	**0.000775**	0.000858	**0.000508**	0.000657
Auto MPG6	**11.7**	11.91	**9.79**	10.56	**0.007361**	0.007428	**0.00558**	0.005621
Auto MPG8	13.64	**10.9**	10.48	**8.6**	0.008476	**0.00734**	0.006421	**0.005208**
Weather Izmir	**3.52**	4.12	**2.99**	3.38	**0.00126**	0.001311	**0.000909**	0.000971
Weather Ankara	**5.33**	6.13	**3.71**	4.054	0.003642	**0.00286**	0.00289	**0.002306**
Abalone	**20.61**	20.72	**16.6**	17.1	0.004836	**0.0042**	**0.00325**	0.003304
California Housing	30.94	**21.94**	25.88	**13.97**	0.002558	**0.00187**	0.00178	**0.0012**
Computer activity	8.84	**2.96**	4.10	**1.86**	0.002581	**0.000446**	0.00131	**0.000304**
Treasury	15.74	**2.16**	12.24	**1.042**	0.00717	**0.00121**	0.0048	**0.000641**
House	44.05	**41.82**	43.36	**25.74**	0.005256	**0.004039**	0.00266	**0.00194**
Mean	15.37	**13.35**	12.125	**9.61**	0.00571	**0.00484**	0.0042	**0.0039**

The average number of the data sets of the minimum error values of FACRM among all evaluation criteria equals 7.5 ((7+9+6+8)/4). The average of mean value of all evaluation criteria are as follows:
FACRM equals 6.876 ((15.37+12.125+0.00571+0.0042)/4) and CART equals 5.742 ((13.35+9.61+0.00484+0.0039)/4).

Table 4-15 presents the number of discovered and diverse rules. FACRM effectively makes a prediction by producing explicit rules, which can be easily understood by the decision makers in the given application domain. Diverse rules are selected based on measuring the distance by using a diversification method, which covers not only the knowledge extracted from high frequency data items, but also knowledge of particular data observations of a low frequency.

CART generates a tree which can be represented as rules, as presented in Table 4-15. The table presents the comparison between the number of rules generated in FACRM and CART. It is clear from the table that our FACRM model generates a fewer number of rules (14.8%) than those generated in CART (85.2%) to give a very close performance. In other words, FACRM produces competitive results with a fewer number of rules compared with CART from the error perspective.

Table 4-15 Number of rules generated in FACRM and CART.

Data set	FACRM		CART
	Number of discovered rules	Number of diverse rules	
Diabetes	**5.4**	0	5.8
Friedman	**114**	2	200.4
Plastic	**10**	0	176.3
Dee	70.2	0.7	**59.4**
Quake	**26.9**	2.2	216.8
Auto MPG6	**58.9**	0	65.3
Auto MPG8	67.7	0	**64.1**
Weather Izmir	**34.2**	0	245.8
Weather Ankara	**15.2**	0	48.2
Abalone	**272**	0.3	524.6
California Housing	**162.9**	3	3135.4
Computer activity	**297.5**	0	1193.9
Treasury	**55.8**	0	188.4
House	**226.1**	3	3533.8
Average	**102**		689.87
85.2%=100% - ((102/689.87)*100%)			

To investigate a statistical significance between FACRM and other models, Wilcoxon rank sum test is used, which is presented in Table 4-16. The comparison between FACRM and ANN shows that the prediction generated by FACRM over ANN is statistically significant for the following data sets: Quake, Auto MPG8, Weather Izmir, California Housing, Computer activity and House. For these data sets the FACRM produces minimum MdAPE. For other data sets (Diabetes, Plastic, Auto MPG6 and Weather Ankara), there is no difference in the prediction by FACRM or ANN. The comparison between FACRM and SVM shows that the prediction generated by FACRM over SVM is statistically significant for these data sets as follows: Plastic, Quake, Auto MPG6, Auto MPG8, Weather Ankara, Computer activity and Treasury. For these data sets the FACRM produces minimum MdAPE. For other data sets (Diabetes, Friedman and Weather Izmir), there is no difference in the prediction by FACRM or SVM. The comparison between FACRM and SR shows that the prediction generated by FACRM over SR is statistically significant for the following data sets: Quake, Auto MPG6, Auto MPG8, Weather Izmir, Weather Ankara, Computer activity, Treasury and House. For these data sets the FACRM produces minimum MdAPE. For other data sets (Diabetes, Friedman and Dee), there is no difference in the prediction by FACRM or SR. The comparison between FACRM and CART shows that the prediction generated by FACRM over CART is statistically significant for the following data sets: Plastic, Quake, Auto MPG6, Weather Izmir, Weather Ankara and Abalone. For these data sets the FACRM produces minimum MdAPE. For other data sets (Diabetes, Friedman and Dee), there is no difference in the prediction by FACRM or CART.

It is worth mentioning that the above analysis demonstrates that FACRM produces satisfactory results compared with other models.

Table 4-16 Statistical significance using Wilcoxon rank sum test.

Data set	FACRM Vs. ANN		FACRM Vs. SVM		FACRM Vs. SR		FACRM Vs. CART	
	Significance		Significance		Significance		Significance	
Diabetes	0.5005	No	0.5628	No	0.5686	No	0.167	No
Friedman	< 0.01**	Yes	0.4547	No	0.1377	No	0.748	No
Plastic	0.1308	No	< 0.01**	Yes	< 0.01**	Yes	0.05*	Yes
Dee	<0.01**	Yes	< 0.01**	Yes	0.2809	No	0.6792	No
Quake	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes	0**	Yes
Auto MPG6	0.3452	No	< 0.01**	Yes	< 0.01**	Yes	0.0191*	Yes
Auto MPG8	0.0303*	Yes	0.0475*	Yes	< 0.01**	Yes	0.01**	Yes
Weather Izmir	< 0.01**	Yes	0.2104	No	< 0.01**	Yes	0.0099**	Yes
Weather Ankara	0.0718	No	< 0.01**	Yes	0.0112*	Yes	0.0128*	Yes
Abalone	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes
California Housing	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes	0**	Yes
Computer activity	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes	0**	Yes
Treasury	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes
House	< 0.01**	Yes	< 0.01**	Yes	< 0.01**	Yes	0**	Yes
*Confidence level % 95	**Confidence level % 99							

4.4.2.3 Second Validation and Comparison Using Nonlinear System Identification

The proposed FACRM prediction model has been validated using well-known benchmark data. It has been used extensively for identification and modelling of ‘gas furnace data’ (Sugeno and Yasukawa, 1993, Sugeno and Tanaka, 1991, Pedrycz, 1984, Tong, 1980, Yinghua and A, 1995, Xu and Lu, 1987, Wang and Langari, 1995, Box and GM, 1970, Kim et al., 1997) which is real life data offered by Box and Jenkins in 1970

(Box and GM, 1970). The data consist of 296 records of single input/output. The input $\mathrm{x(t)}$ represents the gas flow rate into the furnace, while the output $\mathrm{y(t)}$ represents the CO_2 concentration in outlet gas.

FACRM is applied to sustain the identification of CO_2 concentration $\mathrm{y(t)}$. The data set is formed to the common two inputs and one output as $(\mathrm{x(t-4), y(t-1): y(t)})$ which is reduced to 292 records. Figure 4.6 and Figure 4.7 illustrate some characteristic of input $\mathrm{x(t)}$ and output $\mathrm{y(t)}$ of the gas furnace time series data set.

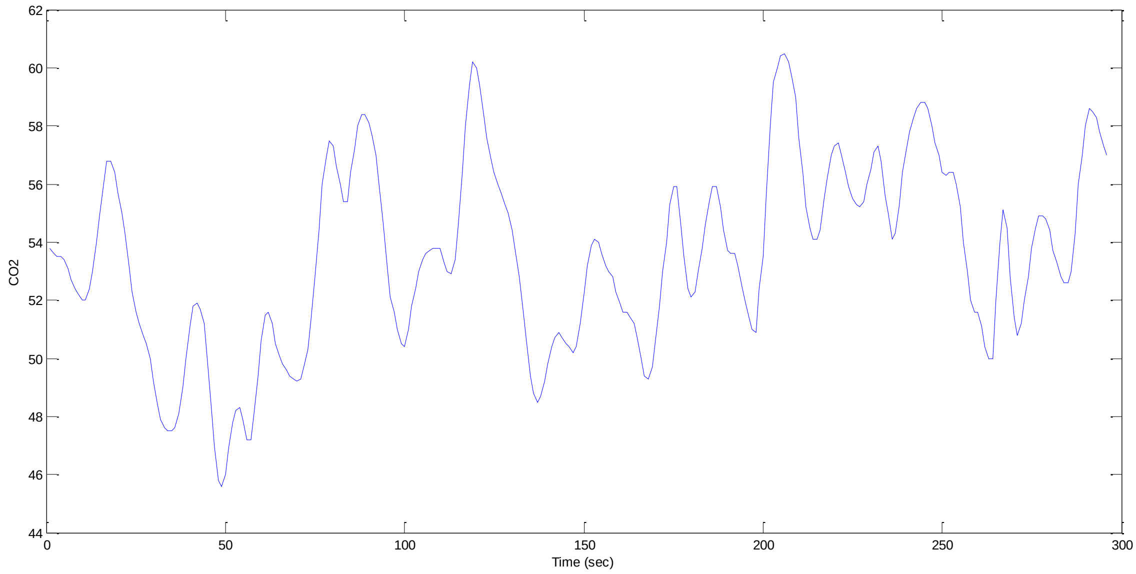


Figure 4.6 The input gas flow rate $\mathbf{x(t)}$.

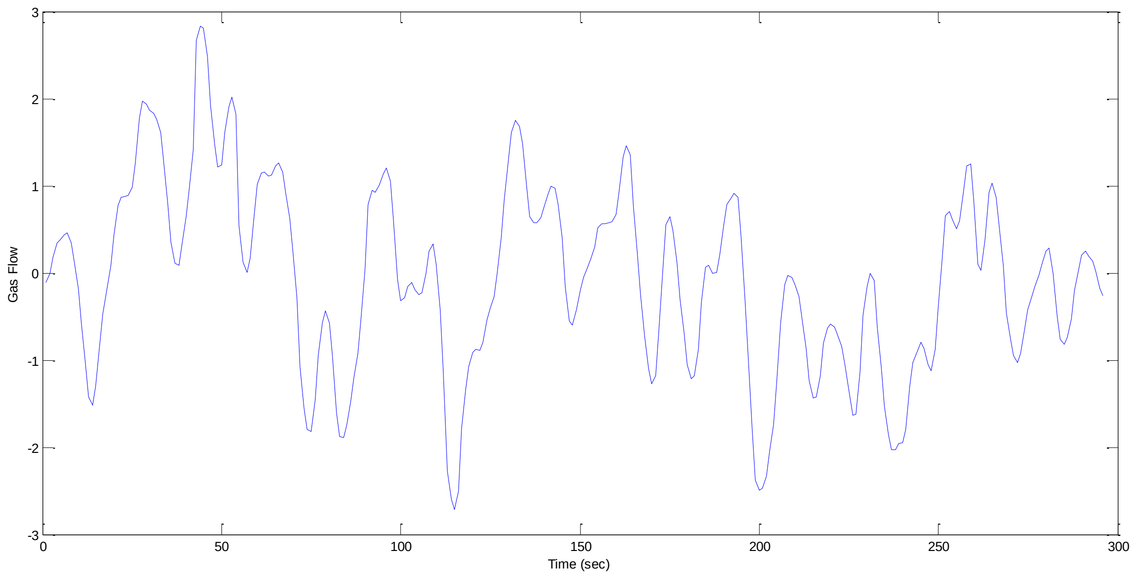


Figure 4.7 The output of CO2 concentration $\mathbf{y(t)}$.

To evaluate the performance of FACRM model, 10-fold cross-validation is used, where LS and *minconf* thresholds are set to 3% and 0.9 respectively. Furthermore, the performance index is the number of rules and Mean Square Error (MSE) of Equation 2-11.

The experimental results show that FACRM has identified the gas furnace prediction with very low MSE equals 0.784×10^{-6} with a small number of rules (6 rules). Also, the results of using min-max normalization are still comparable based on the MSE value and the number of rules. Figure 4.8 shows a slight difference between the actual and predicted values.

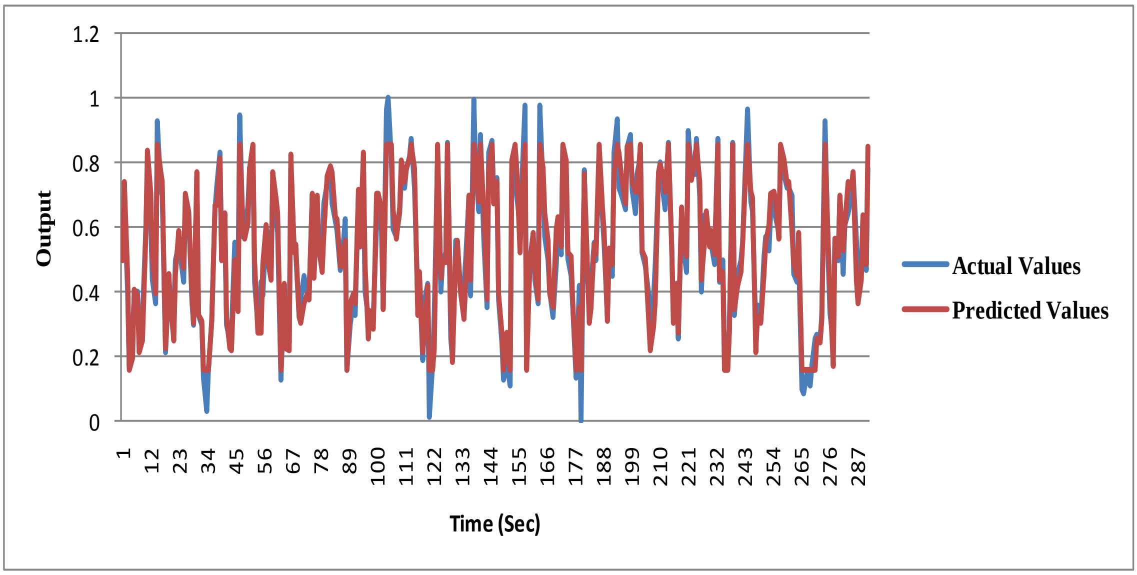


Figure 4.8 The difference between the actual and predicted values.

Table 4-17 summarizes the comparison of the proposed FACRM and other existing models for the identification of the Box-Jenkins problem. The tabulated results show that FACRM is competitive with other models in term of MSE with a fewer number of rules.

From Table 4-17, it can be observed that the proposed model using Euclidian norm normalization produces much better results than those results obtained by other models reported in the literature. On the other hand, the proposed model using min-max normalization produces comparable results with other models. FACRM obtains almost equivalent performance (very close performance) with a fewer number of rules. It is noticed that, the ANFIS, FuNN, FNN-Tool and HyFIS models produce a relatively lower MSE than FACRM, but FACRM generates a fewer number of rules. It is worth mentioning that these models are relatively complex models and difficult to use with high computational time which are based on fuzzy neural network topology, while the proposed model provides the best trade-off between simplicity, result quality and interpretability of the model. In general, the results of the second experiment validation show that the proposed FACRM is a competitive model when compared with other models reported in the literature.

Table 4-17 The comparison of FACRM with different models.

Model in literature	Number of inputs	Number of rules	MSE
Box and Jenkins (ARMA model) (Box and GM, 1970)	5	-	0.71
Xie et al. (Xie et al., 2005)	2	13	0.65151
Tong (Tong, 1980)	2	19	0.469
Lee et al. (Lee et al., 1994)	2	25	0.407
Pedrycz et al (Pedrycz et al., 2002)	2	25	0.395
Sugeno (Sugeno and Yasukawa, 1991)	2	6	0.355
Xu (Xu and Lu, 1987)	2	25	0.328
Pedrycz (Pedrycz, 1984)	2	81	0.32
Lin and Cunningham (Yinghua and A, 1995)	5	4	0.261
Linear model (Sugeno and Yasukawa, 1993)	5	-	0.193
Sugeno and Yasukawa (Position-gradient model) (Sugeno and Yasukawa, 1993)	3	6	0.19
Nie (Nie, 1995)	4	45	0.169
Surmann et al. (Surmann et al., 1993)	2	25	0.160
Hauptmann and Heesche (Hauptmann and Heesche, 1995)	2	10	0.134
Takagi–Sugeno model (Sugeno and Yasukawa, 1993)	6	2	0.068
Wang and Langari (Wang and Langari, 1995)	6	2	0.066
Kim et al. (Kim et al., 1997)	6	2	0.055
Jang (ANFIS model) (Jang et al., 1997)	2	25	0.00073
Kasabov et al. (FuNN model) (Kasabov et al., 1997)	2	7	0.00051
FNN-Tool current model (Almejalli, 2009)	2	15	0.00045
Kim and Kasabov (HyFIS model) (Kim and Kasabov, 1999)	2	15	0.00042
FACRM model (the proposed model using Euclidian Norm)	*2*	*6*	0.784×10^{-6}
FACRM (the proposed model) (Min-Max normalization)	*2*	*6*	*0.0029*

4.5 Summary and Conclusions

In this chapter, an enhanced model for prediction, namely Fuzzy Associative Classification Rule Mining (FACRM), has been proposed and presented. The proposed model is based on the combination of different algorithms and approaches. Firstly, the improved Gustafson-Kessel (G-K) clustering algorithm is applied to transform a crisp quantitative data set into a fuzzy data set. Secondly, Improved Multiple Support Apriori (IMSapriori) algorithm is employed instead of using the global minimum support threshold *minsupp* (single support threshold). IMSapriori improves the classical multiple support approach in order to extract rules including rare terms and to limit insignificant rules. Thirdly, an enhanced method for fuzzy data representation and database scanning format is adapted to improve rules generation process. Finally, a diversification and clustering method is utilized by measuring the distance (dissimilarity) between these rules. Thus, best rules and diverse rules are selected for further prediction reliability.

The performance study of the proposed FACRM model and its capability in discovering significant rules (Fuzzy Rule Base (FRB)) represented by high frequent as well as low frequent data items, have been studied. These extracted rules were used to minimize the prediction error. The experiments were conducted based on two sets of experiments validation. Firstly, the comparison with other prediction models (Artificial Neural Network (ANN), Support Vector Machine (SVM), Stepwise Regression (SR), Classification and Regression Tree (CART)) has been done. Secondly, the identification of a benchmark problem for gas furnace data has been demonstrated. The performance analysis of both methods shows comparable results with respect to reported work in the literature. It can be seen that the proposed FACRM generates a fewer number of rules compared with CART to give approximately a similar performance. The proposed model produces explicit rules for decision makers, which can be seen to have an advantage over the other models (ANN, SVM and SR).

CHAPTER FIVE

5 FEATURE SELECTION METHOD FOR ENHANCING THE PREDICTION MODEL

5.1 Chapter Overview

In this chapter, a new method is proposed for feature (attribute) subset selection entitled Weighting Feature Selection (WFS). The main aim of the proposed method is to enhance the process of data mining techniques (particularly the Fuzzy Associative Classification Rule Mining (FACRM) model developed in Chapter 4) by selecting a suitable number of features. The WFS employs two mechanisms which are weight and intersection operators. The method proposed in this chapter is evaluated using different data sets from the University of California, Irvine (UCI) of machine learning and KEEL (Knowledge Extraction based on Evolutionary Learning) repositories. The experimental results show the significant performance improvement with the proposed method by minimizing the error values and reducing the number of generated rules. Moreover, the WFS provides better results than the result produced by Stepwise Regression (SR).

5.2 Introduction

Data mining techniques have been extensively and widely applied in different domains such as the biomedical (Karabatak and Ince, 2009a, Karabatak and Ince, 2009b, Peng et al., 2010, S´anchez-Monedero et al., 2010) and market basket analysis.

One of the main issues in these domains is high dimensionality data which requires pre-processing. Pre-processing data is a preliminary and essential stage in knowledge discovery and data mining which employs feature selection methods for selecting a subset of features representing the data. Feature selection methods are an important practice in the pre-processing stage (Liu and Yu, 2005). These methods have been applied in many applications, by removing irrelevant and redundant features to assist and enhance model performance and prediction accuracy, i.e. data include irrelevant and redundant features that can cause inadequate predictive accuracy and performance (Peng et al., 2010). Relevant features are quantified as the strength correlation between each feature and the target feature (the output feature/class label feature). Redundant features are quantified as the strength correlation between each feature and other features. Pearson's correlation coefficient is one of the common methods used to measure the correlation (Puuronen et al., 2001).

Feature selection methods/algorithms can be categorized into filter and wrapper models (Bluma and Langley, 1997). The filter model is an essential practice of the data pre-processing stage to select features (feature subset) that have relevance with the target feature (output/class label) based on mutual information and statistical test (Ding and Peng, 2005). In filter model, the feature subsets are rated based on a specific metric, and the data characteristics are examined to find the significant features. The filter model operates independently of any data mining technique and does not work for a specific data mining technique (learning algorithm). It is considered an important stage to reduce data dimension and improve the model accuracy and performance. The advantages of the filter model are represented by the efficiency and independence of the learning technique. The main drawback is that the resultant selected features lead to a different prediction accuracy using a variety of learning techniques. In the wrapper model, different features are evaluated for a specific data mining technique using cross-validation to estimate the usefulness of a feature subset (i.e. prediction accuracy of a feature subset). In other words, this subset is adapted (tailored) for the applied data mining technique. The advantage of the wrapper model is the high probability of prediction accuracy. The common disadvantages are over-fitting and high computational processing. It is noted that the filter model has been observed to be much faster and more practical in large databases than the wrapper model (Williams et al., 2006, Peng et al., 2010, Hall and Smith, 1999, Hall, 2000).

In this chapter, a method called Weighting Feature Selection (WFS) is proposed for feature selection to provide an insight into the influence of feature selection on both prediction error and performance (number of generated rules). The process of careful selection of a suitable number of features can be considered to be data pre-processing. In the context of prediction, feature selection chooses an optimal subset of features among the original features, which can minimize the prediction error and improve performance.

The main focus of this research is to illustrate the benefits of employing the proposed feature selection method as pre-processing, thus enhancing the FACRM model. The prediction error is minimized and the prediction performance is improved by reducing the number of generated rules.

The remaining sections are organized as follows: a literature review and the related work are presented in Section 5.3, a proposed methodology called WFS is illustrated in Section 5.4, experimental results and analysis are discussed in Section 5.5, and finally, the summary and conclusions are presented in Section 5.6.

5.3 Literature Review and Related Work

Correlation-based Feature Selection (CFS) is one of the widespread methods that has been applied for feature selection (Hall and Smith, 1999, Ooi et al., 2007, Lutu and

Engelbrecht, 2010, Williams et al., 2006, Hall, 2000, Hall and Smith, 1998) and is an evaluator method for features subsets. CFS measures the importance and the predictive ability of each feature with the target feature along with the degree of inter-correlation between the features. Specifically, the selected feature subset represents the features of a strong correlation (highly correlated) with target feature (output feature) and a low correlation (a low inter-correlation) between the features. CFS employs a variety of search methods to evaluate feature subsets based on function goodness (Hall and Smith, 1998, Witten and Frank, 2005). The search methods for feature selection aim at finding the best subset of feature from the search space of features. This subset has the capability for better prediction. Several search methods based on heuristic strategy, such as Best First (BF), are often used to find a feature subset in reasonable time (Hall and Smith, 1998). Mainly, the search space can be investigated in both directions: forward selection and backward elimination. For each feature in the forward selection, if it is not included in the current subset then it will be inserted to it to be examined, thus a numeric measure is generated. The impact of inserting the feature is to assess each feature and choose the best one. Conversely, the search is stopped when no enhancement is made by adding the feature to the current subset. This method is called a greedy search which ensures a local instead of a global solution is found. Backward elimination performs in a similar way by removing single feature from the current subset then the evaluation is conducted to find an improvement. Heuristic methods for feature selection are widespread due to their contribution in finding a subset of the selected feature (Anbarasi et al., 2010) as an expensive search is explored to find an optimal subset of features in case of high data dimension (Peng et al., 2005). The search methods of the CFS are demonstrated below (Witten and Frank, 2005). These search methods are also employed in the proposed WFS method.

Best First (BF): a greedy hill climbing search method that accepts the best local modification at each iteration. A local modification to the current feature subset is carried out by either inserting or deleting a single feature. Backtracking is employed after a non-improving iteration. The search method can be progressed forward and backward. Forward starts from the empty set of features, while backward starts from the full set of features or middle point, taking into account all possible single-feature additions and deletions in both search directions.

Greedy Stepwise (GS): a search method performs the same greedy technique as BF search method and has a forward and backward procedure without backtracking. The stopping criteria is applied once the evaluation metric is decreased in case of addition and deletion of the best remaining features.

Exhaustive Search (ES): attempts a sequential search to find a solution by testing all possibilities. It examines the search space of feature subsets initially from the empty set, and then the best subset discovered is reported (revealed).

Random Search (RS): a search method randomly investigates the space of feature subsets. If an initial set is provided, then it seeks subsets that enhance or equal the initial point with the same or fewer number of features.

Genetic Search (GS): a search method utilizes the generic Genetic Algorithm (GA). GA is a search method that can be used for both solving problems and modelling evolutionary systems. Since it is heuristic based (it estimates a solution), similar to most real-life problems where the estimated solution cannot be calculated exactly (Goldberg, 1989). GA is one of the best methods to solve a problem where little information is known, which uses the principles of selection and evolution to produce several solutions to a given problem (Skinner, [accessed October 2010]). A GA starts with a population of individuals (chromosomes) where each individual represents a possible solution for a given problem. Once the genetic representation and the fitness function are defined, GA proceeds to initialize a population of solutions randomly then improve it through repetitive application of crossover, mutation operators. An initial population is specified by determining a list of features and then GA persists in finding an improvement (feasible solution) through a number of generations.

Sequential Feature Selection based on Stepwise Regression (SFS-SR): this has already been discussed in Section 2.13.

Williams et al. (Williams et al., 2006) studied the effect of feature reduction on classification accuracy and computational performance using different machine learning algorithms (Bayesian Network, Naive Bayes Tree and C4.5). The correlation-based feature selection is one of the feature selection methods that has already been used in this study where the results found that the computational performance, such as the classification speed, was more significantly trend effecting than classification accuracy. Also, the results of the classification accuracy using the mentioned algorithms that applied over the selected features are almost similar, but the computational performance is improved.

A performance study of feature selection which applied in the web text classification (categorization) was presented in (Saian and Ku-Mahamud, 2010). An algorithm called Ant-Miner (an optimization algorithm based on an ant colony used for classification problems) (Parpinelli et al., 2002) was used for learning rules to classify web pages. The feature selection method is used to show the performance of the predictive accuracy and number of the generated rules of the selected features. The data set consists of 2,571 web pages. The comparison of the performance results of the Ant-Miner algorithm with the C4.5 algorithm (an algorithm proposed by Quinlan (Quinlan, 1993) based on building a decision tree used for classification problems) was conducted. The random search method of the correlation-based feature selection evaluation was found to provide a higher result in terms of prediction accuracy and number of rules generated using Ant-miner algorithm.

In (Anbarasi et al., 2010) CFS evaluator, in particular, genetic search was applied to find feature subset. Six features were selected out of thirteen features (original feature). The selected feature was later used for classification task. Peng et al. (Peng et al., 2010) developed a novel approach for feature selection in biomedical data ranging between 30 to 50 features where the proposed approach was integrated filter and wrapper models using the sequential search method.

Ding and Peng (Ding and Peng, 2003, Ding and Peng, 2005) developed a method for feature selection called minimum Redundancy Maximum Relevance (mRMR), which used a heuristic algorithm based on Mutual Information (MI). MI is the amount of information that quantifies the relation between a feature F_i and the target feature (output/class). Existence/elimination of a feature F_i will participate in the correct predictive decision. In other words, MI measures the dependency by maximizing it between the selected features and the target feature (Manning et al., 2008 , Peng et al., 2005). mRMR is considered as a filter model that is able to deal with selected features concerning the characteristics of minimum redundancy and maximum relevance criteria. Minimum redundancy reveals that the correlation of the feature set to be minimum to each other, or their mutual Euclidian distance to be maximum. Minimum redundancy reflects the comprehensive representation of the features, for example, in a bioinformatics domain a data set contains 50 features of genes, which contains some genes that are highly correlated, while the number of representative (uncorrelated) genes are few approximately 20 genes, thus, removing 30 genes (highly correlated) from the data will not have an effect on the predictive performance. Therefore, this removing will improve the prediction efficiency. Maximum relevance discloses a high correlation (maximum mutually information) between each feature and the target feature. mRMR is an integration of both criteria relating to minimum redundancy and maximum relevance (Ding and Peng, 2003, Ding and Peng, 2005, Peng et al., 2005). The experimental results were conducted using three classification techniques such as support vector machine, to validate mRMR which was applied on six data sets in gene expression. The results found that the selected features using mRMR have a better accuracy than the selected feature using only maximum relevance criteria selection method (compared with baseline feature set based on standard MI, statistical test that select the ranking m features) (Ding and Peng, 2005).

mRMR is employed in our proposed WFS method due to the following characteristics (Ding and Peng, 2005):

- The most representative feature set (independent features) is selected (minimum redundancy).
- Small number of feature set is found which covers the same space compared with large feature set.
- Integration of maximum relevance and minimum redundancy criteria.

5.4 The Proposed Methodology

The proposed method for feature selection is considered as a filter method, by the integration of different and well-know feature selection methods/algorithms (CFS, SFS-SR and mRMR) to generate the most significant features. Figure 5.1 presents the general overview of the proposed method for feature selection which consists of two main steps:

- Integration of different feature selection methods/algorithms, these are: CFS, SFS-SR and mRMR methods.
- Assessment of different features from different methods/algorithms for selecting the most suitable features using two mechanisms. These

mechanisms are weight and intersection operators as presented in the following definition.

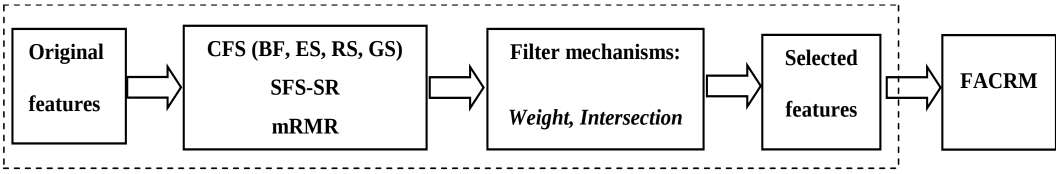


Figure 5.1 The proposed WFS method.

Definition: Let $\mathrm{N} = \{1,2,\dots,\mathrm{n}\}$ be the number of all features and $\mathrm{M} = \{1,2,\dots,\mathrm{m}\}$ be the number of all feature methods. Let $\mathrm{F} = \{\mathrm{F_i} \mid 1 \le \mathrm{i} \le \mathrm{n}\}$ denote a set of features and $\mathrm{MFS} = \{\mathrm{MFS_j} \mid 1 \le \mathrm{j} \le \mathrm{m}\}$ denote the method that is used for features selection. If feature $\mathrm{F_i}$ used by method $\mathrm{MFS_j}$, then $\mathrm{SFGM_j} = \{\mathrm{F_k} \mid \mathrm{k} \in \mathrm{N}\}$ define the selected features generated from method j. Let $\mathrm{SMSF_i}$ be a set of methods that select feature $\mathrm{F_i}$, $\mathrm{SMSF_i} = \{\mathrm{MFS_j} \mid \mathrm{F_i} \in \mathrm{SFGM_j}\}$. $\mathrm{INT_i}$ is the intersection between methods $(\mathrm{MFS_j}, \mathrm{j} \in \mathrm{M})$ for feature $\mathrm{F_i}$ equal zero, if $\mathrm{F_i}$ is not selected at least by one method, otherwise it is equal 1. $\mathrm{UN_i}$ is the union between methods for feature $\mathrm{F_i}$ equal zero, if $\mathrm{F_i}$ is not selected by all methods, otherwise it is equal 1. $\mathrm{W_i}$ be the number of elements in set $\mathrm{SMSF_i}$ and Thw be the threshold weight, where $\mathrm{W_i} \ge$ Thw. This definition of the proposed method is clarified in Figure 5.2.

	$\mathrm{F_1}$	..	$\mathrm{F_n}$
$\mathrm{MFS_1}$	$\mathrm{SFGM_{11}}$	..	$\mathrm{SFGM_{1n}}$
:	:		:
$\mathrm{MFS_m}$	$\mathrm{SFGM_{m1}}$	..	$\mathrm{SFGM_{nm}}$
Intersection	$\mathrm{INT_1}$	..	$\mathrm{INT_n}$
Union	$\mathrm{UN_1}$	..	$\mathrm{UN_n}$
Weight	$\mathrm{W_1}$		$\mathrm{W_n}$

Figure 5.2 The definition of the proposed WFS method.

The research work presented in this chapter aims to improve and evaluate the model prediction (FACRM) in terms of prediction error and performance. The prediction error term is referred to in the common statistical measures presented in Section 2.16, while the term of prediction performance is defined by the number of generated rules. The

WFS is proposed to find the consistent set of features as a filtering model. A flow chart describes the WFS method shown in Figure 5.3.

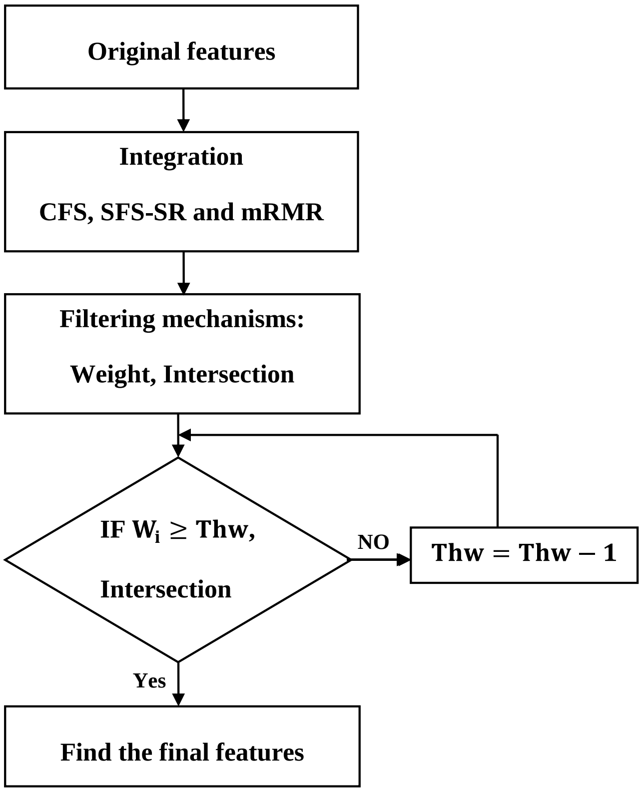


Figure 5.3 Flow chart of WFS method.

Figure 5.4 demonstrates an example of the WFS method. Let $MFS = 5$, $F = 10$, $Thw = 2$. Intersection, union and weight operators are used to filter the selected features extracted from different methods. The features F_5 and F_{10} are filtered and selected using an intersection operator. The weight is computed as follows:

- Counting the number of methods MFS_m that shared the selection of a feature F_i.
- Choosing the feature which has greater than or equal to threshold weight Thw. For example, four features F_2, F_4, F_5, F_{10} are filtered.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
MFS_1	0	0	**1**	**1**	**1**	0	0	0	0	**1**
MFS_2	0	0	0	0	**1**	0	0	0	0	**1**
MFS_3	**1**	**1**	0	0	**1**	0	**1**	**1**	**1**	**1**
MFS_4	0	**1**	0	**1**	**1**	0	0	0	0	**1**
MFS_5	0	0	0	**0**	**1**	0	0	0	0	**1**

Selected features
F3,F4,F5,F10
F5,F10
F1,F2,F5,F7,F8,F9,F10
F2,F4,F5,F10
F5,F10

Intersection	0	0	0	0	**1**	0	0	0	0	**1**
Union	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**
weight	1	**2**	1	**2**	**5**	1	1	1	1	**5**
		F2		**F4**	**F5**					**F10**

Figure 5.4 An example of WFS method.

Figure 5.5 illustrates the procedures of WFS, which selects the features based on two mechanisms; weight and intersection. The union is not used in the experiments any more, as it has produced all features (original data features) in most cases. Thus, the union operator is excluded from the proposed method. mRMR algorithm requires to determine K , where K is the number of selected features. The sensitivity analysis has been conducted to find the best K value. As presented in Figure 5.5, if the number of original features is small (approximately 6 features), then it is found that using weight produces satisfactory results on the tested data sets (as demonstrated in the experiments section). On the other hand, when the number of original features is medium (greater than 6), then the intersection seems to give better results. Alternatively, the weight can be used, since in some data sets, the intersection might not be applicable because at least a feature F_i from all original features could not be selected by all methods.

Input: Original features (all data features).

Output: Filtered features (feature subset).

Method:

1.Find the selected features in each method.

2.Calculate the weight and intersection.

3.Evaluate the selected features as follow:

    For $i = 1$ To n //n is the maximum number of data sets, i is any data set.

        IF NOF between 3 and 6 //NOF is the number of original features.

            Using *Weight* of $K = 2$

        ElseIF NOF between 7 and 21

            Using *Intersection* of $K = 6$ **OR**

            Using *Weight* of $K = 4$

        EndIF

    EndFor

Figure 5.5 The WFS method.

5.5 Experimental Results and Analysis

In this section, the validation of the proposed WFS method for feature selection is conducted. The FACRM model proposed in Chapter 4 is applied in this chapter to assess the selected features (attributes). The main focus of this research is to demonstrate the advantages of the WFS method, which is tested and evaluated using several data sets from the University of California, Irvine (UCI) of machine learning repository (Frank and Asuncion, 2010) and KEEL (Knowledge Extraction based on Evolutionary Learning) repository (Alcalá-Fdez et al., 2011). These data sets were used in Chapter 4 and described in Table 4-6. The common method of 10-fold cross-validation is used where the process of each data set is divided into 10 equal parts of data subsets. Each time, one of the data subset fold is used as testing data and the remaining folds (nine-fold) are used for training of the model. The evaluation measures used to assess the results were also discussed in Chapter 2 (see Section 2.16). Here, the aim of this study is to evaluate the impact of selected features (reduced features) on the prediction error and performance (number of generated rules).

Table 5-1 shows the number of full features (original features) without the target feature (output) in each data set, the number of selected features through WFS, the parameters (LS, *minconf*) used in the experiments chosen based on the sensitivity analysis and finally the number of selected features in SR. The weight value threshold is set to be Thr = 3, this threshold value produced consistent results.

Table 5-1 Full features, selected features in all data sets used in FACRM and SR. The values of LS and *minconf* are also used in FACRM for each data set.

Data set	Full features without target feature	FACRM				SR
		# Selected features	LS		*minconf*	# Selected features
Diabetes	2	2	1	(2.33%)	0.65	2
Friedman	5	4	5	(0.42%)	0.45	4
Plastic	2	2	1	(0.061%)	0.45	**2**
Dee	6	4	20	(5.48%)	0.5	5
Quake	3	1	1	(0.046%)	45	1
Auto MPG6	5	5	15	(3.83%)	0.6	2
Auto MPG8	7	3	1	(0.26%)	0.6	5
Weather Izmir	9	2	155	(10.61%)	0.75	6
Weather Ankara	9	3	10	(2.49%)	0.65	6
Abalone	8	5	200	(4.79%)	0.55	6
California Housing	8	2	100	(0.48%)	0.55	8
Computer activity	21	3	50	(0.61%)	0.7	15
Treasury	15	2	70	(6.7%)	0.8	11
House	16	2	1000	(4.4%)	0.45	15

A comparison is performed between the implementation models of the FACRM, FACRM with WFS and SR. The comparison criteria are based on the error values (error values represented by MAPE, MdAPE, RMSE and MAE) and the number of generated rules. Table 5-2, Figure 5.6, Table 5-3, Table 5-4 and Table 5-5 show overall prediction results on the tested data sets using WFS. The values of MAPE, MdAPE, RMSE and MAE are smaller for the majority of data sets using the WFS method. The results show an enhancement in the quality of FACRM using WFS. Also, the experiments indicate that the FACRM model with the selected feature using the proposed WFS method, is able to produce satisfactory results.

Table 5-2 MAPE (%) values of prediction models for all data sets using full set and selected features.

Data set	FACRM-full set	FACRM-WFS method	SR
Diabetes	**10.63**	**10.63**	12.11
Friedman	21.59	22.64	**19.80**
Plastic	11.1	11.1	**8.85**
Dee	15.17	**13.44**	15.53
Quake	**2.31**	2.59	2.453
Auto MPG6	**11.7**	**11.7**	14.93
Auto MPG8	13.64	**10.97**	16.32
Weather Izmir	3.52	**2.91**	3.872
Weather Ankara	5.33	**4.31**	5.38
Abalone	20.61	20.52	**19.67**
California Housing	**30.94**	31.7	35.66
Computer activity	8.84	**7.19**	11.97
Treasury	15.74	**12.56**	21.91
House	**44.05**	45.01	64.46
Mean	15.37	**14.81**	18.06

Table 5-3 MdAPE (%) values of prediction models in all data using full set and selected features.

Data set	FACRM-full set	FACRM-WFS method	SR
Diabetes	**6.56**	**6.56**	10.38
Friedman	12.95	14.09	**12.53**
Plastic	8.49	8.49	**7.17**
Dee	11.01	**9.68**	12.86
Quake	1.59	**0.94**	2.344
Auto MPG6	**9.79**	**9.79**	12.98
Auto MPG8	10.48	**8.33**	12.9
Weather Izmir	2.99	**2.25**	3.20
Weather Ankara	3.71	**2.98**	3.90
Abalone	16.6	17.05	**16.09**
California Housing	25.88	25.56	**25.14**
Computer activity	**4.10**	4.28	5.145
Treasury	**12.24**	12.40	15.65
House	43.36	**38.80**	45.43
Mean	12.125	**11.51**	13.265

Table 5-4 RMSE values of prediction models for all data sets using full set and selected features.

Data set	FACRM-full set	FACRM-WFS method	SR
Diabetes	**0.0193**	**0.0193**	0.02041
Friedman	0.00537	0.00575	**0.00504**
Plastic	0.003074	0.003074	**0.00245**
Dee	0.008326	**0.00770**	0.008881
Quake	0.000775	0.000822	**0.000677**
Auto MPG6	**0.007361**	**0.007361**	0.007863
Auto MPG8	0.008476	**0.00672**	0.00866
Weather Izmir	0.00126	**0.000914**	0.001291
Weather Ankara	0.003642	**0.00262**	0.00264
Abalone	0.004836	0.00467	**0.00358**
California Housing	0.002558	**0.00254**	0.00266
Computer activity	0.002581	0.00141	**0.00124**
Treasury	0.00717	**0.00368**	0.008351
House	0.005256	0.00482	**0.004137**
Mean	0.00571	**0.00509**	0.00556

Table 5-5 MAE values of prediction models for all data sets using full set and selected features.

Data set	FACRM-full set	FACRM-WFS method	SR
Diabetes	**0.01526**	**0.01526**	0.01708
Friedman	0.00427	0.00456	**0.00390**
Plastic	0.0025	0.0025	**0.001974**
Dee	0.00669	**0.00606**	0.006774
Quake	**0.000508**	0.00057	0.000532
Auto MPG6	**0.00558**	**0.00558**	0.00587
Auto MPG8	0.006421	**0.00502**	0.00659
Weather Izmir	0.000909	**0.00070**	0.000781
Weather Ankara	0.00289	**0.00200**	0.00204
Abalone	0.00325	0.0033	**0.00256**
California Housing	0.00178	0.0018	**0.00171**
Computer activity	0.00131	0.000936	**0.000779**
Treasury	0.0048	**0.00324**	0.00507
House	0.00266	**0.00228**	0.002316
Mean	0.0042	**0.00384**	0.00414
Difference(RMSE-MAE)	0.00151	0.00125	0.00142

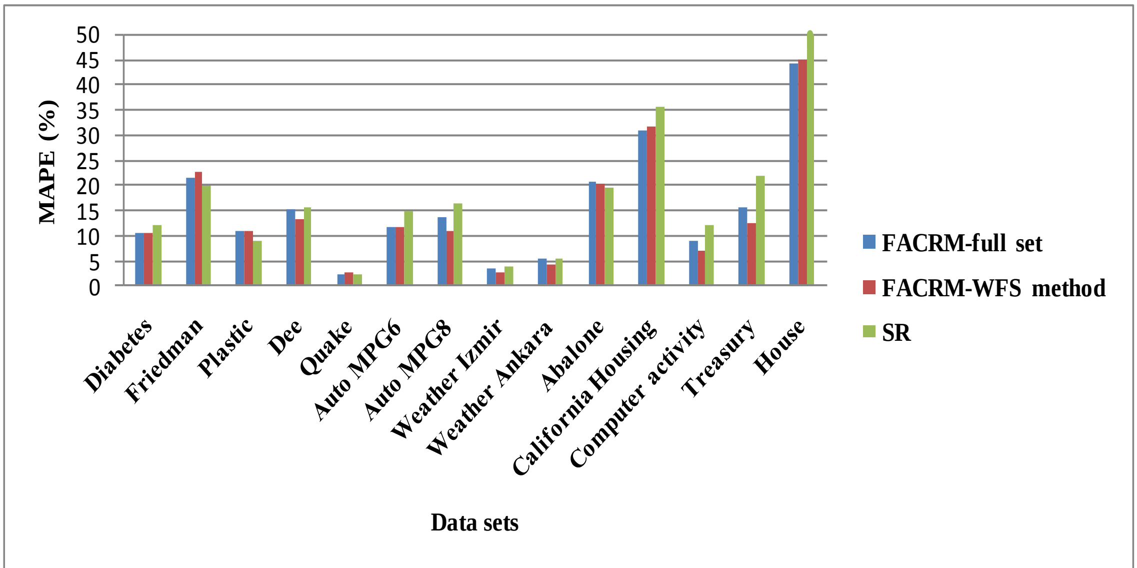


Figure 5.6 The comparison between FACRM-full set, FACRM-WFS method and SR using MAPE values for all data sets.

The results of the comparisons of FACRM, FACRM with WFS and SR presented in Table 5-2, Figure 5.6, Table 5-3, Table 5-4 and Table 5-5 confirmed a number of observations. Firstly, using FACRM with WFS produced better results on the tested data sets than using full data features (original features). The second observation is that using FACRM with WFS provided better results than the results achieved by SR.

As mentioned in Section 4.4.2.2 the importance of using MAE and RMSE is to detect the variation in the prediction error values. The difference between RMSE and

MAE in FACRM-WFS equals 0.00125, as illustrated in Table 5-5, which reduces the difference compared with FACRM- full set (0.00151). This means that FACRM-WFS is accurate for most cases (individual error values of the prediction), while there are some large error values that affect on RMSE. FACRM-WFS has a lower difference value than SR in terms of the difference between RMSE and MAE. Also, FACRM-WFS produced the best results in terms of MdAPE (the minimum mean value of MdAPE for all data sets as shown in Table 5-3). This means that, FACRM-WFS is relatively an accurate and consistent model.

Furthermore, Table 5-6 expresses the number of generated rules by utilizing WFS. As expected, the results reveal that the number of rules decreases as the number of data features decreases. This means that the selected features by the proposed WFS method have a better prediction performance.

Table 5-6 The discovered and diverse rules generated by FACRM using full set and selected features.

Data set	Full set of features		Using WFS	
	# Discovered rules	# Diverse rules	# Discovered rules	# Diverse rules
Diabetes	**5.4**	0	**5.4**	0
Friedman	114	2	**78.2**	1.6
Plastic	**10**	0	**10**	0
Dee	70.2	0.7	**34.9**	0
Quake	26.9	2.2	**4**	0
Auto MPG6	**58.9**	0	**58.9**	0
Auto MPG8	67.7	0	**27.9**	1.5
Weather Izmir	34.2	0	**9**	0
Weather Ankara	15.2	0	**7.2**	0
Abalone	272	0.3	**40.1**	0.9
California Housing	162.9	3	**11**	1.9
Computer activity	297.5	0	**19.7**	2
Treasury	55.8	0	**11**	0
House	226.1	3	**12.1**	0
Mean	102		**24.1**	

The key findings of this study are:

- The proposed method is able to deal with several data sets of different domains. Thus, the identification and selection of suitable features is performed to improve the prediction model.

- The prediction error is not greatly improved (i.e. the error value remains same or is slightly decreased) in some data sets. However the prediction performance is enhanced by speeding up prediction and minimizing the number of generated rules.
- The sensitivity analysis of the selected features is accomplished to express the role and importance of the proposed method.

Assessing the results presented in Table 5-2 indicates that the MAPE value of the Friedman data set selected 4 features (approximately the same as the original feature) among all features (5 features) and is equal to 22.64. One can notice that the MAPE value increased from 21.59 to 22.64 when using the WFS method. This method slightly deteriorates the accuracy of the prediction model by increasing the error value. The MAPE of the Treasury data set selected 2 features (very low number of features) out of 15 features and is equal to 12.56. As can be seen in Table 5-2 the MAPE is decreased from 15.74 to 12.56 by applying WFS. Therefore, this method impacts the prediction model by decreasing the error value. This can be explained as follows:

- Pearson’s correlation coefficient measure is calculated between the original features and the target feature to find the relevant features, especially for the Friedman and Treasury data sets. The coefficient values of the selected features for the Friedman data set are 0.433, 0.371, 0.615 and 0.275, respectively. In the Treasury data set, the coefficient values are 0.992 and 0.995, respectively. It is observed that in the Friedman data set the correlation values are not as strongly correlated compared with the values in the Treasury data set which are very strongly correlated. This implies a reduction of the error value.
- Dimension reduction and feature selection methods have been considered and used to reduce high dimensional data. Features discarding is

designed for features that do not contribute much to model performance and facilitation of model interpretation. For instance, it seems that 4 features selected in the Friedman and 2 features selected in the Treasury data sets are already a small number of selected features. Therefore, all features may be required for a good model performance (even a small increase or decrease in the error value).

- Feature selection methods can discard important information (some features). These methods remove features which are not correlated with the target features (feature relevance), but show a low degree of inter-correlation among them (redundant feature). In other words, the feature subset is under-described. This will affect the model performance.

5.6 Summary and Conclusion

In this chapter, a new method is proposed namely Weighting Feature Selection (WFS) method. The proposed method is simple, flexible and reliable, employing two mechanisms for selecting a suitable number of features. These mechanisms are weight and intersection operators. The WFS provides a filtering approach as data pre-processing with the aim of improving the prediction performance by minimizing the error rate and reducing the number of generated rules.

Feature selection method is one of the most significant aspects in the knowledge discovery process. One of the major problems of the knowledge discovery models is use of high dimensional data. In this case, increasing dimensionality of data will effect negatively the model's performance and interpretability. The proposed WFS method for feature selection was found to perform effectively by robustly selecting a small number of features (feature subset). This subset is able to perform better on prediction tasks than a larger number of features (reduced noise). Experiments on several benchmark data sets showed that this method is practical, much more robust and better than the results produced by Stepwise Regression (SR). Also, the results clearly confirm the great potential of the proposed method for enhancing the model's prediction in most cases.

CHAPTER SIX

6 CONCLUSIONS AND FUTURE WORK

6.1 Observations and Conclusions

The research work presented in this thesis has investigated the design of a model for predicting a future value accurately. A new model related to a knowledge discovery process for predicting a future value was proposed. Extensive research has been conducted to build a prediction model in a particular domain. This study was focused on building a model to be applied in different benchmark data sets in a wide range of application domains. In addition, this model has been developed to tackle the problem of extracting association rules (useful knowledge) from a quantitative data. The future value prediction using a combination and adaptation of fuzzy clustering, multiple support thresholds and associative classification approaches were investigated. The performance of the proposed model was shown by conducting a set of experiments on benchmark data sets. Finally, a feature selection method entitled Weighting Feature Selection (WFS) was proposed. WFS considers as a data pre-processing, in order to improve the performance of the proposed prediction model by minimizing the error rate and reducing the number of generated rules.

The *first stage* of the proposed Knowledge Discovery (KD) prediction model design presented in Chapter 3 includes the following steps: (1) quantitative data is transformed into fuzzy data by using Fuzzy C-Means (FCM) algorithm; (2) Fuzzy Association Rules (FARs) are extracted from fuzzy data by applying Apriori approach;

(3) FARs are filtered and stored in a Knowledge Base (KB); (4) Fuzzy Inference System (FIS) is used to command the KB for prediction in a particular application domain.

In this context, FARs mining was applied based on the combination of the FCM algorithm and Apriori approach. Subsequently, the Apriori approach was utilized to extract FARs in the form of “IF-Then” from fuzzy data. Next, those rules are filtered by considering only the rules where their consequent parts (“Then” parts) included a target attribute (dependent/output attribute), and then the filtered FARs are used for building a KB.

Finally, FIS was used to command the extracted FARs that are stored in the KB. As a result, the prediction model in the first stage has been evaluated and validated using two case studies of different data set sizes in a road traffic management domain. This road traffic data has been generated using a traffic simulation model called the METANET.

The experimental results using benchmark evaluation measures for error estimation have been used to compare two data sets of different size. One of the conclusions resulting from the experiments is that the data set of a large size produced better results with lower MAPE, NMAE, NRMSE and Uncorr than the results produced in the data set of a small size.

FCM is one of the fuzzy clustering algorithms, which was employed in order to handle a quantitative data since association rule mining, in particular Apriori algorithm, fails to deal with quantitative data directly. Quantitative data is essential and important in the sense that it includes distinct values, which requires an efficient and simple techniques to transform it into fuzzy data. The result of FCM is to transform the quantitative data into fuzzy data.

The *second stage* of this research, described in Chapter 3, was to improve the prediction model (KD model) of the first stage. The main advantages of the prediction model presented in the second stage are: (1) handling unbalance data through using multiple support thresholds of Minimum Item Support (MIS) instead of using one single support *minsupp* threshold; and (2) finding the best and most diverse rules (representative rules) through proposing a method for rules diversification.

The proposal in this study was to incorporate a diversification method as a supporting method within the prediction model. The advantage of this diversification method is to find the best and representative rules, for covering hidden and infrequent knowledge space. This enhances the ability of the prediction model to produce rules to cover the data set used in the model learning process. Hence, the prediction model has extracted useful knowledge (high quality and diverse rules) for predicting new input data cases of such data set.

The prediction model in the second stage includes the following steps: (1) FCM was used to transform quantitative data into fuzzy ones; (2) Multiple Support Apriori (MSapriori) approach was employed for extracting FARs from fuzzy data; (3) a method for rules diversification was proposed in order to select the best and representative rules. The diversification method has utilized the sharing function technique used in multi-objective optimization for clustering FARs. The representative rules are a small number of rules that cover data set cases, especially those cases of a low frequency in a particular data set; (4) the significant and representative FARs are stored in a KB; (5) FIS is used to command the KB for prediction in a particular application domain.

The model in this stage has been applied in two case studies; the first case study was applied to a benchmark data set called Abalone, while the second one was applied to a data set in road traffic domain. This model has produced satisfactory results with lower MAPE as compared to other prediction approach as reported in the literature.

It was observed that, the generated rules are highly dependent on the threshold values (such as Least Support (LS) and β values) used in the proposed prediction model. Thus, the selection of an appropriate threshold value leads to the selection of a suitable number of rules.

The experimental results show that the proposed prediction model can be applied to a wider range of application domains. The knowledge base contains diverse, useful knowledge and a reasonable number of rules. Therefore, the proposed model is reliable and helps the user/human expert easily to understand an application domain.

It was noticed that employing a multiple support thresholds approach improved the results of the proposed prediction model. The main idea underlying this approach was to deal with an unbalanced data set and to tackle the problem of rare items, which aims to generate frequent patterns with rare items (attributes). The diversification method was an essential phase to select the best (significant) and representative rules, and then the representative rules are found by clustering the rules to ensure the diversity in the knowledge base. The significant and diverse rules make the prediction model more generic and reliable. Although using diverse rules is important to make the prediction model robust/generic, in some instances it might result in a slight increase in the prediction error value. It is observed from the experimentation that the best and diverse rules contribute to building a generic prediction model in a particular application domain to predict a future value for new input data case accurately.

The *third stage* of this research study presented in Chapter 4, further investigations have been carried out to improve and develop the prediction model of the second stage. The outcome of the third stage is a new novel hybrid model for predicting a future value, namely Fuzzy Associative Classification Rule Mining (FACRM). This model adapts recent algorithms/approaches: (1) the improved Gustafson-Kessel (G-K) fuzzy clustering algorithm has been used to transform quantitative data into fuzzy data; (2) the improved multiple support algorithm has been utilized to extract the frequent patterns that contain the rare items (attributes) and to limit the combinatorial explosion (uninteresting frequent itemsets); (3) the vertical data format scanning for fuzzy data has been adapted to improve the efficiency of the rules extraction process; (4) the associative classification approach has been employed for rules pruning directly; (5) the proposed diversification method used in the second stage was also applied in this stage.

The hybrid model was tested on different benchmark data sets (14 data sets) from the University of California, Irvine (UCI) of machine learning and Knowledge Extraction based on Evolutionary Learning (KEEL) repositories. The effectiveness and performance of this model have been evaluated and validated using the following two experimentations with the common 10-fold cross-validation method.

In the first experimentation, an empirical performance study has been conducted by comparing the proposed hybrid model with common and well-known existing prediction models, namely (Artificial Neural Network (ANN), Support Vector Machine (SVM), Stepwise Regression (SR) and Classification and Regression Tree (CART)). The common statistical measures were used in order to evaluate the results such as MAPE, MdAPE, RMSE and MAE. Also, Wilcoxon rank sum test was used to assess the statistical significance between FACRM and other models. The results showed that the proposed FACRM produced better results for the majority of the data sets among all prediction models. In terms of the number of generated rules, the results showed that, FACRM produces less number of rules in comparison with CART rule base model.

In the second validation experimentation, a well-known benchmark problem called the Box-Jenkins problem of gas furnace data was used. This data has been used extensively for identification and modelling. The experiments were performed using two normalization methods for trustworthy comparison to other existing and well-known models in the literature. Furthermore, a comparative analysis demonstrated that FACRM is competitive with these exiting models in terms of MSE and produced fewer number of rules.

The proposed hybrid model features the following: (1) Fuzzy Classification Association Rules (FCARs) (or Fuzzy Rule Base (FRB)) discovered from fuzzy data directly, and improves the model's performance. In contrast with the FARs extraction process, FARs requires a filtration technique to select only rules that contain a target attribute in their consequent parts (right-hand side of the rules); (2) useful knowledge (rules) are extracted, i.e. extracting frequent patterns that contain rare items and avoid non-dominating rules; (3) the best and representative rules are determined and selected for reliable prediction.

FCM was used in Chapter 3 while G-K algorithm was applied in Chapters 4 and 5. It is worth mentioning that based on the test data sets and existing literature, both the FCM and G-K algorithm provide approximately the same results for determining the fuzzy sets.

The limitations of the proposed model can be summarized as follows:

- The proposed prediction model has been applied to a quantitative data set (numeric data).
- The prediction accuracy deteriorates approximately in case of high dimensional (large data size) and outliers data.
- High dimensional data and correlated data attributes have generated a large number of rules. This makes the prediction slightly slow by effecting the efficiency (performance and computational time) of FIS.
- Tuning parameters and thresholds (such as *minsupp*, *minconf*, β value) have been necessary to check the prediction error value. Appropriate parameters and thresholds have to be selected based on a sensitivity analysis for each data set.

Finally, towards performance enhancement of the proposed FACRM model, a new method has been proposed for feature selection (attribute selection) introduced in

Chapter 5, namely Weighting Feature Selection (WFS). The proposed method is based on two mechanisms; the mechanisms are weight and intersection operators.

It is observed that, the proposed WFS method performed effectively by robustly selecting a small number of features (feature subset). This subset is able to perform better on prediction tasks than a larger number of features by reducing noise.

The results obtained have demonstrated that WFS is practical and offers a potential for enhancing the model prediction in comparison to FACRM without WFS. A comparative analysis has demonstrated that using WFS as a filtering model (data pre-processing) implies better results than Stepwise Regression (SR).

Regarding the effect of WFS on improving the prediction performance, it was found that WFS can reduce the prediction error value and minimize the number of generated rules. However, in some cases applied to some data sets the WFS degraded the prediction performance by increasing the prediction error value. Nevertheless, the proposed method produced satisfactory results by decreasing the prediction error value among overall data sets.

In summary, an investigation into several approaches and methods to build and develop an effective and reliable model for prediction was proposed and experimented in this thesis. Further investigation was carried out to develop a method for feature selection in order to improve and enhance the proposed prediction model.

6.2 Future Work

Some suggestions for future work that can be built from the research outcomes presented in this thesis are as follows:

- In the proposed prediction model in Chapter 3 and Chapter 4, it was assumed that the maximum cluster number is set to be four clusters (four fuzzy sets were used for each attribute) to reduce the complexity of the prediction model and FIS. It would be interesting to enhance the proposed prediction model through the clustering validation measure. Several methods have been proposed in the literature to measure validity of the number of fuzzy clustering. Xie and Beni (XB) (Xie and Beni, 1991) is one recommended clustering validity measure that could be used to verify the number of fuzzy clustering in the improved G-K or FCM algorithms. The minimum value of XB is the optimal clusters number.

- The current clustering algorithms are not always able to find the optimal fuzzy sets. Thus, further enhancement for the current prediction model can be gained by using a dynamic adjustment and tuning the membership functions of the fuzzy sets, which can be achieved using Genetic Algorithm (GA). The importance of the rules (knowledge) quality depends on the superiority of the membership functions, with the purpose of reflecting more accuracy on the discovered knowledge. Improving the performance of the membership function is achieved by adjusting (tuning) the fuzzy set, which optimizes the range of the fuzzy sets to find a suitable membership function (Kaya and Alhajj, 2003, Hong et al., 2010, Hong et al., 2006, Chen et al., 2008a, Chen et al., 2008b).
- In a supervised learning approach, the target attribute (output attribute) is determined as a special attribute called label. In case the label value is discrete or category, it is called the class label and hence the task applied is called classification. Otherwise, if the label value is continuous, then the task applied is called regression. The proposed prediction model in this thesis employs a supervised learning approach using the target attribute (label) of continuous values. It is worth investigating the

performance of applying the proposed prediction model to a classification problem with some modifications. The modifications can be accomplished by developing an effective method for ranking the extracted rules in order to build a classifier model. A classifier model can be constructed by two phases. In the first phase, the rules are generated by utilizing Fuzzy Associative Classification Rules (FACR) algorithm presented in Section 4.4.1. In the second phase, the rules are ranked and evaluated by developing an effective method to find out the representative (coverage) rules from a training data (Thabtah et al., 2005, Pach et al., 2008).

- The proposed Weighting Feature Selection (WFS) method has been experimented for data sets with up to 22 features. However, further performance analysis of the WFS method would be interesting using extensive experiments for very high dimensional data sets with a very high number of features. In addition, the proposed prediction model has been applied to complete data sets but not for the data set with missing values. It would be interesting to investigate the proposed model for some data sets including missing values.
- It would be interesting to further investigate the performance of computational cost, which can be improved by incorporating and adapting a Frequent Pattern Growth (FP-Growth) approach. This approach stores the database in a condensed form (see Section 2.3.3), in order to improve the candidate itemsets generation process.
- The experimental results carried out on this thesis have shown the effectiveness and reliability of the proposed prediction model in several data sets of different application domains. Further studies can be

conducted by investigating the performance of the proposed prediction model in other wider problem domains. In addition, the proposed feature selection method and prediction model in this thesis facilitate KB building process, which can provide a basis for a Decision Support System (DSS) for an application area.

References

Abonyi, J., Feil, B. and Abraham, A. (2005) Computational Intelligence in Data Mining. *Informatica,* **29,** 3-12.

Abu-Nimeh, S., Nappa, D., Wang, X. and Nair, S. (2007) A comparison of machine learning techniques for phishing detection. In: *In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit.* Pittsburgh, Pennsylvania,USA: ACM, Vol. 269 pp. 60-69.

Agarwal, R., Aggarwal, C. and Prasad, V. (2000) Depth first generation of long patterns. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.*

Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.* Washington, D.C., United States, pp. 207-216.

Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules in large databases. In: *Proceeding of 20th International Conference on Very Large Databases (VLDB).* Santiago, Chile, pp. 487-499.

Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L. and Herrera., F. (2011) KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing,* **17** (2-3)**,** 255-287. [accessed July 2010].

Almejalli, K. (2009). *An Intelligent Real-Time Decision Support System for Road Traffic Management*. PhD Thesis. University of Bradford.

Anbarasi, M., Anupriya, E. and Iyengar, N. (2010) Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology* **2**(10)**,** 5370-5376

Antonie, M. and Zaïane, O. (2002) Text document categorization by term association. In: *Proceedings of the IEEE International Conference on Data Mining.* Maebashi City, Japan, pp. 19-26.

Antonie, M., Zaïane, O. and Coman, A. (2003) Associative classifiers for medical images. In: *Mining Multimedia and Complex Data.* LNAI, Springer Berlin / Heidelberg, Vol. 2797, pp. 68-83.

Arauzo-Azofra, A., Aznarte, J. L. and Benítez, J. M. (2011) Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems With Applications,* **38** (7)**,** 8170-8177.

Aziz, S. A. and Parthiban, J. ([accessed December 2010]) Fuzzy Logic Available from: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/sbaa/report.html.

Azzeh, M., Neagu, D. and Cowling, P. (2010) Fuzzy grey relational analysis for software effort estimation. *Empirical Software Engineering*, **15** (1)**,** 60-90.

Azzeh, M., Neagu, D. and Cowling, P. I. (2011) Analogy-based software effort estimation using Fuzzy numbers. *Journal of Systems and Software*, **84** (2)**,** 270-284.

Babuska, R., Van Der Veen, P. and Kaymak, U. (2002) Improved covariance estimation for Gustafson-Kessel clustering. In: *IEEE International Conference on Fuzzy Systems.* Honolulu, Hawaii, pp. 1081-1085.

Balasko, B., Abonyi, J. and Feil, B. (2008) Fuzzy clustering and data analysis toolbox Department of Process Engineering, University of Veszprem, Hungary, Available from: http://www.fmt.vein.hu/softcomp/fclusttoolbox/.

Bezdek, J. C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms* Kluwer Academic Publishers Norwell, MA, USA.

Bluma, A. L. and Langley, P. (1997) Selection of relevant features and examples in machine learning. *Artificial intelligence*, **97** (1-2)**,** 245-271.

Bose, I. and Mahapatra, R. K. (2001) Business data mining—a machine learning perspective. *Information & Management*, **39** (3)**,** 211-225.

Box, G. and Gm, J. (1970) Time Series Analysis Forecasting and Control. Holden-Day, San Francisco.

Breiman, L. (1984) *Classification and regression trees.* Chapman & Hall/CRC.

Chang, C.-C. and Lin, C.-J. (2011) LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2** (3)**,** 1-27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. [accessed July 2010].

Chang Chien, Y. W. and Chen, Y. L. (2010) Mining associative classification rules with stock trading data-A GA-based method. *Knowledge-Based Systems*, **23** (6)**,** 605-614.

Chapple, M. ([accessd February 2008]) Regression. Available from: http://databases.about.com/od/datamining/g/regression.htm.

Charles, C. M. (1995) *Introduction to educational research (2 nd edition).* Longman.

Chen, C. H., Hong, T. P. and Tseng, V. S. (2006a) A Cluster-Based Fuzzy-Genetic Mining Approach for Association Rules and Membership Functions. In: *The 2006 IEEE International Conference on Fuzzy Systems.* Vancouver, BC, Canada, pp. 1411-1416.

Chen, C. H., Hong, T. P. and Tseng, V. S. (2008a) A cluster-based genetic-fuzzy mining approach for items with multiple minimum supports. In: *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD'08).* Springer-Verlag, pp. 864-869.

Chen, C. H., Tseng, V. S. and Hong, T. P. (2008b) Cluster-based evaluation in fuzzy-genetic data mining. *Fuzzy Systems, IEEE Transactions on*, **16** (1)**,** 249-262.

Chen, G., Liu, H., Yu, L., Wei, Q. and Zhang, X. (2006b) A new approach to classification based on association rule mining. *Decision Support Systems*, **42** (2)**,** 674-689.

Cheng, M.-Y. and Roy, A. F. V. (2010) Evolutionary fuzzy decision model for cash flow prediction using time-dependent support vector machines. *International Journal of Project Management,* **29** (1)**,** 56-65.

Cho, S., Kim, J. and Bae, J. (2009) An integrative model with subject weight based on neural network learning for bankruptcy prediction. *Expert Systems with Applications,* **36** (1)**,** 403-410.

Chu, M. T., Shyu, J., Tzeng, G. H. and Khosla, R. (2007) Comparison among three analytical methods for knowledge communities group-decision analysis. *Expert Systems With Applications,* **33** (4)**,** 1011-1024.

Cohen, W. W. (1995) Fast effective rule induction. In: *Proceedings of the International Conference on Machine Learning.* pp. 115–123.

Collis, J. and Hussey, R. (2003) *Business research: a practical guide for undergraduate and postgraduate students.* Palgrave Macmillan.

Creswell, J. W. (2003) *Research Design: Qualitative and Quantitative approaches (2nd edition).* Sage.

Creswell, J. W. (2005) *Educational research: Planning, conducting, and evaluating quantitative and qualitative research (2nd edition).* Prentice Hall.

Dahou, Z., Mehdi Sbarta , Z., Castel, A. and Ghomari, F. (2009) Artificial neural network model for steel-concrete bond prediction. *Engineering Structures,* **31** (8)**,** 1724-1733.

Deb, K. (2001) *Multi-objective optimization using evolutionary algorithms.* Wiley.

Dechang, P. and Xiaolin, Q. (2008) A New Fuzzy Clustering Algorithm on Association Rules for Knowledge Management. *Information Technology Journal,* **7** (1)**,** 119-124.

Delgado, M., Marin, N., Martin-Bautista, M. J., Sanchez, D. and Vila, M. A. (2003) Mining Fuzzy Association Rules: An Overview. In: *Proceedings Of the BISC International Workshop on Soft Computing for Internet and Bioinformatics.* Springer, Vol. 224, pp. 351-374.

Ding, C. and Peng, H. (2003) Minimum redundancy feature selection from microarray gene expression data. In: *Proceedings of 2nd IEEE Computational Systems Bioinformatics Conference (CSB 2003).* Stanford, CA: IEEE, pp. 523-528.

Ding, C. and Peng, H. (2005) Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics & Computational Biology,* **3** (2)**,** 185-205.

Duan, J., Wang, W., Zeng, J., Zhang, D. and Shi, B. (2009) A prediction algorithm for time series based on adaptive model selection. *Expert Systems with Applications,* **36** (2)**,** 1308-1314.

Dunham, M. H. (2002) *Data Mining: Introductory and Advanced Topics.* Prentice Hall PTR Upper Saddle River, NJ, USA.

Dunkel, B. and Soparkar, N. (1999) Data organization and access for efficient data mining. In: *Proceedings of the 15th International Conference on Data Engineering (ICDE'99).* Sydney, Australia, p. 522.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996a) From data mining to knowledge discovery in databases. *AI magazine,* **17** (3)**,** 37-54.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996b) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* **39** (11)**,** 27-34.

Fayyad, U. M. (1996) Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert,* **11** (5)**,** 20-25.

Frank, A. and Asuncion, A. (2010) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences. [accessed July 2010].

Fu, A. W., Wong, M. H., Sze, S. C., Wong, W. C., Wong, W. L. and Yu, W. K. (1998) Finding Fuzzy Sets for the Mining of Fuzzy Association Rules for Numerical Attributes. In: *in Proc. Int. Symp. Intelligent Data Engineering Learning (IDEAL'98).* Hong Kong, pp. 263–268.

Gedikli, F. and Jannach, D. (2010) Neighborhood-restricted mining and weighted application of association rules for recommenders. In: *the 11th International Conference on Web Information Systems Engineering (WISE 2010).* Hong Kong, China: Springer Berlin / Heidelberg, Vol. 6488, pp. 157-165.

Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Grivas, G. and Chaloulakou, A. (2006) Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric environment,* **40** (7)**,** 1216-1229.

Guillaume, S. (2001) Designing fuzzy inference systems from data: an interpretability-oriented review. *Fuzzy Systems, IEEE Transactions on,* **9** (3)**,** 426-443.

Hahsler, M., Grun, B. and Hornik, K. (2005) arules–A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software,* **14** (15)**,** 1-25.

Hall, M. A. (2000) Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: *Proceedings of the Seventeenth International Conference on Machine Learning.* Morgan Kaufmann Publishers Inc., pp. 359-366.

Hall, M. A. and Smith, L. A. (1998) Practical feature subset selection for machine learning. In: *in Proceedings 21st Australian Computer Science Conference.* pp. 181–191.

Hall, M. A. and Smith, L. A. (1999) Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In: *In Proceedings of the Florida Articial Intel ligence Symposium (FLAIRS-99).* AAAI Press, pp. 235-239.

Han, J. and Fu, Y. (1995) Discovery of Multiple-Level Association Rules from Large Databases. In: *In Proc. of the 21st Int'l Conference on Very Large Databases.* Zurich, Switzerland.

Han, J., Pei, J. and Yin, Y. (2000) Mining Frequent Patterns Without Candidate Generation. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data.* Dallas, Texas, United States: ACM, pp. 1–12.

Han, S., Schneider, S. M. and Evans, R. G. (2003) Evaluating Cokriging for Improving Soil Nutrient Sampling Efficiency. *Transaction-American Society of Agricultural Engineers,* **46** (3)**,** 845-850.

Hand, D., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining.* Cambridge: MIT Press.

Hauptmann, W. and Heesche, K. (1995) A neural net topology for bidirectional fuzzy-neuro transformation. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE/IFES).* Yokohama: IEEE, Vol. 3, pp. 1511-1518.

Hong, T. P., Chen, C. H. and Tseng, V. S. (2010) Genetic-Fuzzy Data Mining Techniques. In: *2010 IEEE International Conference on Granular Computing.* San Jose, California: IEEE, pp. 26-27.

Hong, T. P., Chen, C. H., Wu, Y. L. and Lee, Y. C. (2006) A GA-based Fuzzy Mining Approach to Achieve a Trade-off Between Number of Rules and Suitability of Membership Functions. *Soft Computing-A Fusion of Foundations, Methodologies and Applications,* **10** (11)**,** 1091-1101.

Hong, T. P., Kuo, C. S. and Wang, S. L. (2004) A fuzzy AprioriTid mining algorithm with reduced computational time. *Applied Soft Computing Journal,* **5** (1)**,** 1-10.

Hong, T. P., Lin, K. Y. and Chien, B. C. (2003b) Mining Fuzzy Multiple-Level Association Rules from Quantitative Data. *Applied Intelligence,* **18** (1)**,** 79-90.

Hong, T. P., Lin, K. Y. and Wang, S. L. (2003a) Fuzzy data mining for interesting generalized association rules. *Fuzzy Sets and Systems,* **138** (2)**,** 255-269.

Hu, Y. H. and Chen, Y. L. (2006) Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decision Support Systems,* **42** (1)**,** 1-24.

Huang, M. J., Tsou, Y. L. and Lee, S. C. (2006) Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. *Knowledge-Based Systems,* **19** (6)**,** 396-403.

Imberman, S. P. (2002) Effective Use of the KDD Process and Data Mining for Computer Performance Professionals. Computer Measurement Group; 1997, Vol. 2, pp. 611-620.

Ivanciuc, O. (2007) Applications of support vector machines in chemistry. *Reviews in Computational Chemistry,* **23,** 291-400.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) Data clustering: a review. *ACM computing surveys (CSUR),* **31** (3)**,** 264-323.

Jang, J., Sun, C. and Mizutani, E. (1997) *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence.* Prentice Hall.

Janssens, D., Wets, G., Brijs, T. and Vanhoof, K. (2003) Integrating Classification and Association Rules by proposing adaptations to the CBA Algorithm. In: *Proceedings of the 10th International Conference on Recent Advances in Retailing and Services Science.* Portland, Oregon, USA.

Kanellopoulos, Y., Makris, C. and Tjortjis, C. (2007) An improved methodology on information distillation by mining program source code. *Data & Knowledge Engineering,* **61** (2)**,** 359-383.

Kannan, S. and Bhaskaran, R. (2009) Association Rule Pruning based on Interestingness Measures with Clustering. *International Journal of Computer Science Issues(IJCSI),* **6** (1)**,** 35-43.

Kannan, S. and Bhaskaran, R. (2010) Role of Interestingness Measures in CAR Rule Ordering for Associative Classifier: An Empirical Approach. *Journal of Computing,* **2** (1)**,** 8-15.

Kannan, S. R. and Genova, I. (2005) Segmentation Of MRI Using New Unsupervised Fuzzy C Mean Algorithm. *ICGST-GVIP* **5, Issue** (2).

Karabatak, M. and Ince, M. (2009a) An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications,* **36** (2)**,** 3465-3469.

Karabatak, M. and Ince, M. (2009b) A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications,* **36** (10)**,** 12500-12505.

Kasabov, N., Kim, J., Watts, M. and Gray, A. (1997) FuNN/2--A fuzzy neural network architecture for adaptive learning and knowledge acquisition. *Information Sciences,* **101** (3-4)**,** 155-175.

Kaya, M. and Alhajj, R. (2003) Facilitating fuzzy association rules mining by using multi-objective genetic algorithms for automated clustering. In: *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03).* pp. 561-564.

Khan, M., Muyeba, M. and Coenen, F. (2008) Mining Fuzzy Association Rules from Composite Items. *Artificial Intelligence in Theory and Practice II,* **276,** 67-76.

Kianmehr, K. and Alhajj, R. (2008) CARSVM: A class association rule-based classification framework and its application to gene expression data. *Artificial Intelligence in Medicine,* **44** (1)**,** 7-25.

Kim, E., Park, M., Ji, S. and Park, M. (1997) A new approach to fuzzy modeling. *Fuzzy Systems, IEEE Transactions on,* **5** (3)**,** 328-337.

Kim, J. and Kasabov, N. (1999) HyFIS: adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems. *Neural Networks,* **12** (9)**,** 1301-1319.

Kiran, R. and Reddy, P. (2009) An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. In: *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM.* Nashville, TN, USA, pp. 340-347.

Kiran, R. and Reddy, P. (2010a) Mining Rare Association Rules in the Datasets with Widely Varying Items' Frequencies. In: *Proceedings of 15th International Conference on Database Systems for Advanced Applications (DASFAA 2010)* Tsukuba, Japan: Springer, pp. 49-62.

Kiran, R. U. and Reddy, P. K. (2010b) Improved approaches to mine rare association rules in transactional databases. In: *Proceedings of the Fourth SIGMOD PhD Workshop on Innovative Database Research.* Indianapolis, Indiana, USA: ACM, pp. 19-24.

Kryszkiewicz, M. (1998) Representative Association Rules. In: *Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge*

Discovery and Data Mining PAKDD-98. Springer-Verlag, Vol. 1394, pp. 198-209.

Kryszkiewicz, M. (2009) Closures of downward closed representations of frequent patterns. In: *Proceedings of the 4th International Conference on Hybrid Artificial Intelligence Systems HAIS 2009.* Salamanca, Spain: Springer-Verlag, Vol. 5572, pp. 104-112.

Kryszkiewicz, M. and Rybinski, H. (1999) Incomplete database issues for representative association rules. *Proceedings of the 11th International Symposium on Foundations of Intelligent Systems.* Springer-Verlag.

Kumar, R. (2010) *Research Methodology: A Step-by-Step Guide for Beginners.* Sage

Kumar, R., Jayaraman, V. and Kulkarni, B. (2005) An SVM classifier incorporating simultaneous noise reduction and feature selection: illustrative case examples. *Pattern Recognition,* **38** (1)**,** 41-49.

Kuok, C. M., Fu, A. and Wong, M. H. (1998) Mining fuzzy association rules in databases. *ACM SIGMOD Record,* **27** (1)**,** 41-46.

Lancaster, G. (2005) *Research Methods in Management: A concise introduction to research in management and business consultancy.* Butterworth-Heinemann.

Le, T. T. N., Huynh, H. X. and Guillet, F. (2009) Finding the Most Interesting Association Rules by Aggregating Objective Interestingness Measures. In: *Knowledge Acquisition: Approaches, Algorithms and Applications: Pacific Rim Knowledge Acquisition Workshop, PKAW 2008, Hanoi, Vietnam, December 15-16, 2008, Revised Selected Papers.* Springer-Verlag, pp. 40-49.

Lee, Y., Hwang, E. and Shih, Y. (1994) A combined approach to fuzzy model identification. *Systems, Man and Cybernetics, IEEE Transactions on,* **24** (5)**,** 736-744.

Lei, Z. and Ren-Hou, L. (2007) An Algorithm for Mining Fuzzy Association Rules Based on Immune Principles. In: *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007.* Boston, MA, pp. 1285-1289.

Lenca, P., Meyer, P., Vaillant, B. and Lallich, S. (2008) On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research,* **184** (2)**,** 610-626.

Leopold, E. and Kindermann, J. (2002) Text categorization with support vector machines. How to represent texts in input space? *Machine Learning,* **46** (1)**,** 423-444.

Lesot, M.-J. and Kruse, R. (2006) Gustafson-Kessel-like clustering algorithm based on typicality degrees. *International Conference on Information Processing and Management of Uncertainty.* Paris, France.

Liu, B., Hsu, W. and Ma, Y. (1998) Integrating Classification and Association Rule Mining. In: *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD'98).* pp. 80-86.

Liu, B., Hsu, W. and Ma, Y. (1999) Mining association rules with multiple minimum supports. In: *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99).* San Diego, California, United States: ACM, pp. 337-341.

Liu, H., Jeng, B., Yih, J. and Yu, Y. (2009) Fuzzy C-Means Algorithm Based on Standard Mahalanobis Distances. In: *Proceedings of the 2009 International Symposium on Information Processing (ISIP'09).* Huangshan, P. R. China, pp. 422-427.

Liu, H., Wu, D., Yih, J. and Liu, S. (2008a) Fuzzy possibility C-mean based on complete mahalanobis distance and separable criterion. In: *Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition.* Hong Kong: IEEE, pp. 50-55.

Liu, H. and Yu, L. (2005) Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering,* **17** (4)**,** 491-502.

Liu, Y. Z., Jiang, Y. C., Liu, X. and Yang, S. L. (2008b) CSMC: A combination strategy for multi-class classification based on multiple association rules. *Knowledge-Based Systems,* **21** (8)**,** 786-793.

Liyanage, C., Elhag, T., Ballal, T. and Li, Q. (2009) Knowledge communication and translation–a knowledge transfer model. *Journal of Knowledge Management,* **13** (3)**,** 118-131.

Lu, J., Xu, B. and Jiang, J. (2003a) A prediction method of fuzzy association rules. In: *IEEE International Conference on Information Reuse and Integration.*, pp. 98-103.

Lu, J., Xu, B. and Yang, H. (2003b) A classification method of fuzzy association rules. In: *Proceedings of the Second IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications.* Lviv, Ukraine, pp. 248-251.

Lutu, P. E. N. and Engelbrecht, A. P. (2010) A decision rule-based method for feature selection in predictive data mining. *Expert Systems with Applications,* **37** (1)**,** 602-609.

Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval.* (http://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html) Cambridge, UK: Cambridge University Press.

Marukatat, R. (2006) Structure-Based Rule Selection Framework for Association Rule Mining of Traffic Accident Data. In: *Proceeding in International Conference on Computational Intelligence and Security.* Guangzhou, pp. 781-784.

Matteucci, M. ([accessed Novmber 2008]) A Tutorial on Clustering Algorithms: Fuzzy C-Means Clustering. Available from: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html.

Mendes, E., Di Martino, S., Ferrucci, F. and Gravino, C. (2007) Effort estimation: how valuable is it for a web company to use a cross-company data set, compared to using its own single-company data set? In: *Proceedings of the 16th international conference on World Wide Web.* Banff, Alberta, Canada: ACM, pp. 963-972.

Mendes, E., Watson, I., Triggs, C., Mosley, N. and Counsell, S. (2003) A comparative study of cost estimation models for web hypermedia applications. *Empirical Software Engineering,* **8** (2)**,** 163-196.

Menon, R., Tong, L. H. and Sathiyakeerthi, S. (2005) Analyzing textual databases using data mining to enable fast product development processes. *Reliability Engineering and System Safety,* **88** (2)**,** 171-180.

Messmer, I. A. (2007) A simulation program for motorway networks. *METANET, Technical University of Crete.*

Mitra, S., Pal, S. K. and Mitra, P. (2002) Data mining in soft computing framework: A survey. *Neural Networks, IEEE Transactions on,* **13** (1)**,** 3-14.

Myers, M. (1997) Qualitative research in information systems. *Management Information Systems Quarterly,* **21** (2)**,** 241-242, MISQ Discovery, archival version, June 1997, http://www.misq.org/discovery/MISQD_isworld/. MISQ Discovery, updated version, last modified: August 10, 2011 www.qual.auckland.ac.nz.

Negnevitsky, M. (2005) *Artificial Intelligence: A Guide to Intelligent Systems.* Addison-Wesley.

Nie, J. (1995) Constructing fuzzy model by self-organizing counterpropagation network. *IEEE Transactions on Systems, Man and Cybernetics,* **25** (6)**,** 963-970.

Ooi, C. H., Chetty, M. and Teng, S. W. (2007) Differential prioritization in feature selection and classifier aggregation for multiclass microarray datasets. *Data Mining and Knowledge Discovery,* **14** (3)**,** 329-366.

Özbakir, L., Baykasoglu, A. and Kulluk, S. (2010) A soft computing-based approach for integrated training and rule extraction from artificial neural networks: DIFACONN-miner. *Applied Soft Computing,* **10** (1)**,** 304-317.

Özbakir, L., Baykasoglu, A., Kulluk, S. and Yapici, H. (2009) TACO-miner: An ant colony based algorithm for rule extraction from trained neural networks. *Expert Systems with Applications,* **36** (10)**,** 12295-12305.

Pach, F. P., Gyenesei, A. and Abonyi, J. (2008) Compact fuzzy association rule-based classifier. *Expert Systems with Applications,* **34** (4)**,** 2406-2416.

Park, J., Chen, M. and Yu, P. (1995) An effective hash-based algorithm for mining association rules. *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data.* San Jose, California, United States.

Parpinelli, R. S., Lopes, H. S. and Freitas, A. A. (2002) Data mining with an ant colony optimization algorithm. *IEEE Transactions on Evolutionary Computation,* **6** (4)**,** 321-332.

Pedrycz, W. (1984) An identification algorithm in fuzzy relational systems. *Fuzzy Sets and systems,* **13** (2)**,** 153-167.

Pedrycz, W., Lam, P. and Rocha, A. (2002) Distributed fuzzy system modeling. *IEEE Transactions on Systems, Man and Cybernetics,* **25** (5)**,** 769-780.

Peng, H., Long, F. and Ding, C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **27** (8)**,** 1226-1238.

Peng, Y., Wu, Z. and Jiang, J. (2010) A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics,* **43** (1)**,** 15-23.

Polydoras, G., Anagnostopoulos, J. and Bergeles, G. (1998) Air quality predictions: dispersion model vs Box-Jenkins stochastic models. An implementation and comparison for Athens, Greece. *Applied Thermal Engineering,* **18** (11)**,** 1037-1048.

Put, R., Perrin, C., Questier, F., Coomans, D., Massart, D. and Vander Heyden, Y. (2003) Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship studies. *Journal of Chromatography A,* **988** (2)**,** 261-276.

Puuronen, S., Tsymbal, A. and Skrypnyk, I. (2001) Correlation-Based and Contextual Merit-Based Ensemble Feature Selection. In: *Proceedings of 4th International Conference on Advances in Intelligent Data Analysis (IDA 2001).* Cascais, Portugal: Springer-Verlag, Vol. 2189, pp. 135-144.

Quek, C., Pasquier, M. and Lim, B. B. S. (2006) POP-TRAFFIC: a novel fuzzy neural approach to road traffic analysis and prediction. *Intelligent Transportation Systems, IEEE Transactions on,* **7** (2)**,** 133-146.

Quinlan, J. R. (1993) *C4. 5: programs for machine learning.* Morgan Kaufmann.

Remenyi, D., Williams, B., Money, A. and Swartz, E. (1998) *Doing Research in Business and Management: An Introduction to Process and Method.* Sage.

Romero, C. and Ventura, S. (2007) Educational data mining: A survey from 1995 to 2005. *Expert Systems With Applications,* **33** (1)**,** 135-146.

Russel, S. J. and Norvig, P. (2003) *Artificial intelligence: A Modern Approach.* Prentice-Hall.

S´Anchez-Monedero, J., Cruz-Ram´ırez, M., Fern´Andez-Navarro, F., Fern´Andez, J. C., Guti´Errez, P. A. and Herv´as-Mart´ınez, C. (2010) On the suitability of Extreme Learning Machine for gene classification using feature selection. In: *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA'10).* Cairo, Egypt: IEEE.

Saian, R. and Ku-Mahamud, K. R. (2010) Comparison of Attribute Selection Methods for Web Texts Categorization. In: *Second International Conference on Computer and Network Technology (ICCNT)* Bangkok IEEE, pp. 115-118.

Sapankevych, N. and Sankar, R. (2009) Time series prediction using support vector machines: a survey. *Computational Intelligence Magazine, IEEE,* **4** (2)**,** 24-38.

Saunders, M. N. K., Thornhill, A. and Lewis, P. (2007) *Research methods for business students.* Prentice Hall.

Shaw, M. J., Subramaniam, C., Tan, G. W. and Welge, M. E. (2001) Knowledge management and data mining for marketing. *Decision Support Systems,* **31** (1)**,** 127-137.

Shepperd, M. and Kadoda, G. (2001) Comparing software prediction techniques using simulation. *IEEE Transactions on Software Engineering,* **27** (11)**,** 1014-1022.

Shihab, A. I. and Burger, P. (1998) The Analysis of Cardiac Velocity MR Images Using Fuzzy Clustering. In: *Proceeding of SPIE Medical Imaging 1998—Physiology and Function from Multidimensional Images.* Vol. 3337, pp. 176–183.

Shin, K., Lee, T. and Kim, H. (2005) An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications,* **28** (1)**,** 127-135.

Shitong, W., Chung, F. L. and Hongbin, S. (2005) Fuzzy taxonomy, quantitative database and mining generalized association rules. *Intelligent Data Analysis,* **9** (2)**,** 207-217.

Simon, H. (1999) *Neural networks: a comprehensive foundation.* Prentice hall.

Skinner, M. ([accessed October 2010]) Genetic Algorithms Overview. Available from: http://geneticalgorithms.ai-depot.com/Tutorial/Overview.html.

Sousa, S., Martins, F., Alvim-Ferraz, M. and Pereira, M. (2007) Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software,* **22** (1)**,** 97-103.

Srikant, R. and Agrawal, R. (1996) Mining quantitative association rules in large relational tables. In: *Proceedings of the 1996 ACM SIGMOD international conference on Management of data.* Montreal, Quebec, Canada pp. 1-12.

Stathacopoulou, R., Magoulas, G. D., Grigoriadou, M. and Samarakou, M. (2005) Neuro-fuzzy knowledge processing in intelligent learning environments for improved student diagnosis. *Information Sciences,* **170** (2-4)**,** 273-307.

Sugeno, M. and Tanaka, K. (1991) Successive identification of a fuzzy model and its applications to prediction of a complex system. *Fuzzy Sets and systems,* **42** (3)**,** 315-334.

Sugeno, M. and Yasukawa, T. (1991) Linguistic modeling based on numerical data. In: *Proceedings IFSA '91.* Brussels: Vol. 91, pp. 264–267.

Sugeno, M. and Yasukawa, T. (1993) A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on fuzzy systems,* **1** (1)**,** 7-31.

Surmann, H., Kanstein, A. and Goser, K. (1993) Self-organizing and genetic algorithms for an automatic design of fuzzy control and decision systems. In: *The First European Congress on Fuzzy and Intelligent Technologies, EUFIT' 93.* Aachen, pp. 1097-1104.

Suzuki, E. (2009) Compression-Based Measures for Mining Interesting Rules. In: *Proceedings of the 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: Next-Generation Applied Intelligence.* Tainan, Taiwan: Springer-Verlag Vol. 5579, pp. 741-746.

Tan, P.-N., Steinbach, M. and Kumar, V. (2006) *Introduction to Data Mining.* Addison-Wesley.

Thabtah, F. (2007) A review of associative classification mining. *The Knowledge Engineering Review,* **22** (1)**,** 37-65.

Thabtah, F., Cowling, P. and Hammoud, S. (2006) Improving rule sorting, predictive accuracy and training time in associative classification. *Expert Systems with Applications,* **31** (2)**,** 414-426.

Thabtah, F., Cowling, P. and Peng, Y. (2005) MCAR: multi-class classification based on association rule. In: *Proceeding of the Third IEEE International Conference on Computer Systems and Applications.* Cairo, Egypt, pp. 1-7.

Tong, R. (1980) The evaluation of fuzzy models derived from experimental data. *Fuzzy Sets and systems,* **4** (1)**,** 1-12.

Trochim, W. M. K. ([accessed September 2011]) Research Methods Knowledge Base, 2006. Available from: http://www.socialresearchmethods.net/kb/index.php.

Uday Kiran, R. and Reddy, K. (2009) An improved frequent pattern-growth approach to discover rare association rules. In: *International Conference on Knowledge Discovery and Information Retrieval.* pp. 43-52.

Vo, B. and Le, B. (2009) A Novel Classification Algorithm Based on Association Rules Mining. In: *Knowledge Acquisition: Approaches, Algorithms and Applications Pacific Rim Knowledge Acquisition Workshop, PKAW 2008.* Hanoi, Vietnam: Springer, Heidelberg, Vol. LNCS, pp. 61-75.

Wang, L. and Langari, R. (1995) Building Sugeno-type models using fuzzy discretization and orthogonal parameter estimation techniques. *IEEE Transactions on Fuzzy Systems,* **3,** 454–458.

Wang, W., Men, C. and Lu, W. (2008) Online prediction model based on support vector machine. *Neurocomputing,* **71** (4-6)**,** 550-558.

Williams, N., Zander, S. and Armitage, G. (2006) A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review,* **36** (5)**,** 5-16.

Willmott, C. J. and Matsuura, K. (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research,* **30** (1)**,** 79-82.

Witten, I. H. and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann Publishers.

Wong, B., Bodnovich, T. and Selvi, Y. (1997) Neural network applications in business: a review and analysis of the literature (1988-1995). *Decision Support Systems,* **19** (4)**,** 301-320.

Wu, G. and Lo, S. (2010) Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network. *Expert Systems with Applications,* **37** (7)**,** 4974-4983.

Xie, K., Chen, Z. and Qiu, Y. (2005) Fuzzy Forecast Modeling for Gas Furnace Based on Fuzzy Sets and Rough Sets Theory. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing,* **3642,** 614-623.

Xie, X. and Beni, G. (1991) A validity measure for fuzzy clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* **13** (8)**,** 841-847.

Xu, B., Lu, J., Zhang, Y., Xu, L., Chen, H. and Yang, H. (2003) Parallel algorithm for mining fuzzy association rules. In: *In Proceeding of the International Conference on Cyberworlds.* pp. 288-293.

Xu, C. and Lu, Y. (1987) Fuzzy model identification and self-learning for dynamic systems. *Systems, Man and Cybernetics, IEEE Transactions on,* **17** (4)**,** 683-689.

Xu, R. and Wunsch, D. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks,* **16** (3)**,** 645-678.

Ye, Y. and Chiang, C. C. (2006) A Parallel Apriori Algorithm for Frequent Itemsets Mining. In: *Proceedings of the Fourth International Conference on Software Engineering Research,Management and Applications IEEE.* Washington, USA, pp. 87-94.

Yeh, I. (1998) Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research,* **28** (12)**,** 1797-1808.

Yeh, I. (2007) Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites,* **29** (6)**,** 474-480.

Yin, X. and Han, J. (2003) CPAR: Classification based on Predictive Association Rules. In: *Proceedings of the third SIAM International Conference on Data Mining.* San Francisco, CA, USA, pp. 331-335.

Yinghua, L. and A, C. G. (1995) A new approach to fuzzy-neural system modeling. *IEEE Transactions on fuzzy systems,* **3** (2)**,** 190-198.

Zadeh, L. A. (1965) Fuzzy sets *Information and Control,* **8** (3)**,** 338-353.

Zadeh, L. A. (1988) Fuzzy logic. *Computer,* **21** (4)**,** 83-93.

Zaki, M. (2000) Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on,* **12** (3)**,** 372-390.

Zaki, M., Parthasarathy, S., Ogihara, M. and Li, W. (1997) New algorithms for fast discovery of association rules. In: *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining.*

Zaki, M. J. and Gouda, K. (2003) Fast vertical mining using diffsets. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* Washington, D.C., ACM.

Zhang, L., Shi, Y. and Yang, X. (2005) A Fuzzy Mining Algorithm for Association-Rule Knowledge Discovery. In: *Proceedings of the Eleventh Americas Conference on Information Systems.* Omaha, NE, USA.

Zhang, M. and He, C. (2010) Survey on Association Rules Mining Algorithms. *Advancing Computing, Communication, Control and Management***,** 111-118.

Zhang, W. (1999) Mining Fuzzy Quantitative Association Rules. In: *in Proc. IEEE Int. Conf. Tools with Artificial Intelligence.* Chicago, pp. 99–102.

Zhang, X., Chen, G. and Wei, Q. (2009) Building a highly-compact and accurate associative classifier. *Applied Intelligence***,** 1-13.