AN IMPROVED IMPUTATION METHOD BASED ON FUZZY C-MEANS AND PARTICLE SWARM OPTIMIZATION FOR TREATING MISSING DATA

NURUL ASHIKIN BINTI SAMAT

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

AN IMPROVED IMPUTATION METHOD BASED ON FUZZY C-MEANS AND PARTICLE SWARM OPTIMIZATION FOR MISSING DATA

NURUL ASHIKIN BINTI SAMAT

A thesis submitted in
fulfillment of the requirement for the award of the
Degree of Master of Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

AUGUST, 2017

To

Mak, Ayah,

&

sister & brothers

ACKNOWLEDGEMENT

Alhamdullilah, all praises to Allah. I would like to express my thanks to those who helped me with various aspects in conducting research and the writing of this thesis. First and foremost, to my supervisor, Assoc. Prof. Dr. Mohd Najib Mohd Salleh who sparked my interest in my research area, encouraged and guided me to complete this journey with his immense knowledge. His advice, patience, support and motivation helped me to make it through the ups and downs of this process.

I would like to express sincere gratitude to Malaysian Higher Education (MOHE) for providing sponsorship for study fees under the MyBrain15 program. I also would like to sincerely acknowledge the Universiti Tun Hussein Onn Malaysia (UTHM) for providing financial assistance under Postgraduate Incentive Research Grant (GIPS). Both financial supports motivate me to complete my study. I would like to gratefully thank to the Centre for Graduate Studies (CGS) and the Research, Innovation, Commercialization, and Consultancy Office (ORICC) for their hospitality during my study. I also wish to say thank to Faculty of Computer Science and Information Technology Postgraduate friends and staff who had supported my research work in one way or another.

Finally, I am forever indebted to my parents, brothers, and sister for their unbounded support and encouragement. My father's great advice, my mother's continuous prayers, and their love and dedication provided the foundations for this work. I love them to the moon and back.

ABSTRACT

Data mining techniques are used in various industries, including database marketing, web analysis, information retrieval and bioinformatics to gain a better knowledge extraction. However, if data mining techniques are applied on real datasets, a problem that often comes up is that missing values occur in the datasets. Since the missing values may confuse the data mining process and causing the knowledge extracted unreliable, there is a need to handle the missing values. Therefore, researchers are coming out with imputation methods in the preprocessing stage. Although there are many imputation methods such as Mean, *k*-Nearest Neighbor (*k*-NN) and Fuzzy C-Means are implemented by other researchers, accuracy for the replace values is still in infancy. In this study, an imputation based on FCM and Particle Swarm Optimization (PSO) has been developed to get better imputation values. FCM has ability to cluster the data into two or more subsets with the different membership values and gives better coverage to find the correlation between the dataset. While, PSO is a swarm optimization algorithm that effectively find the optimum imputation values with less parameters to adjust. Then, FCMPSO was trained with seven artificial missing ratios from 1% to 30% for Cleveland Heart Disease dataset and real missing values in Framingham Heart Disease dataset to get the complete dataset. Then, the complete dataset was trained with Decision Tree algorithm to observe the performance in terms of accuracy. The FCMPSO results gives a better RMSE value for 30% missing ratios with 0.0237 compared to Mean, *k*-NN, and FCM with 0.0250, 0.0402 and 0.0249 respectively. Next, the analysis of proposed imputation on classification accuracy shows an improvement with 81.67% for Cleveland Heart Disease and 86.3% for Framingham Heart Disease compared to other imputation methods. Based on the results, the imputation values are slightly accurate compared to other imputation methods and therefore, increased the accuracy of Decision Tree classification.

ABSTRAK

Teknik perlombongan data digunakan di dalam pelbagai industi bagi mendapatkan pengetahuan yang lebih baik. Walaubagaimanapun, masalah data hilang selalu terjadi di dalam data sebenar. Data yang hilang boleh mengelirukan proses perlombongan data dan menyebabkan pengetahuan yang diekstrak tidak dapat dipercayai. Oleh itu, terdapat kepentingan untuk mengendalikan data yang hilang. Para penyelidik, telah mengaplikasikan kaedah imputasi di dalam fasa preproses. Walaupun terdapat banyak kaedah imputasi seperti kaedah Min, k-Nearest Neighbor (k-NN) dan Fuzzy C-Means (FCM) yang dilaksanakan oleh penyelidik lain, ketepatan untuk nilai ganti masih boleh diperbaiki. Dalam kajian ini, satu kaedah imputasi berdasarkan FCM dan Particle Swarm Optimization (PSO) telah dibangunkan bagi mendapatkan nilai imputasi yang lebih baik. FCM mempunyai keupayaan untuk mengumpulkan data ke dalam dua atau lebih kumpulan dengan nilai keahlian yang berlainan serta memberikan liputan yang lebih baik untuk mencari hubungan di antara dataset. Sementara itu, PSO adalah algoritma pengoptimumam yang baik bagi mencari nilai imputasi yang optimum dengan parameter yang sedikit untuk diubah suai. Kemudian, FCMPSO telah diuji dengan tujuh nisbah data hilang dari 1% hingga 30% untuk dataset Penyakit Jantung Cleveland dan nilai sebenar yang hilang dalam dataset Penyakit Jantung Framingham untuk mendapatkan dataset lengkap. Kemudian, dataset lengkap dilatih dengan algoritma Keputusan Pohon untuk melihat prestasi dari segi ketepatan. Keputusan FCMPSO memberikan nilai RMSE yang lebih baik untuk 30% nisbah hilang dengan 0.0237 berbanding Mean, k-NN, dan FCM masing-masing dengan 0.0250, 0.0402 dan 0.0249. Seterusnya, bagi ketepatan klasifikasi menunjukkan peningkatan sebanyak 81.67% untuk Penyakit Jantung Cleveland dan 86.3% untuk Penyakit Jantung Framingham berbanding kaedah imputasi yang lain. Berdasarkan hasilnya, nilai imputasi lebih tepat dibandingkan dengan kaedah imputasi lain dan oleh itu, meningkatkan ketepatan pengklasifikasian Pokok Keputusan.

TABLE OF CONTENTS

DECLARATION ii

DEDICATION iii

ACKNOWLEDGEMENT iv

ABSTRACT v

ABSTRAK vi

TABLE OF CONTENTS vii

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ALGORITHMS xiii

LIST OF SYMBOLS AND ABBREVIATIONS xiv

LIST OF APPENDICES xv

LIST OF PUBLICATIONS xvi

LIST OF AWARD xvii

CHAPTER 1 INTRODUCTION 1

1.1 Introduction 1

1.2 Problem Statement 3

1.3 Aim of Study 4

1.4 Objectives of the Study 4

1.5 Scope of Study 5

1.6 Significance of Study 5

1.7 Thesis Outline 5

CHAPTER 2 LITERATURE REVIEW 7

2.1 Introduction 7

2.2 An Overview of Data Mining 8

2.3 Concept of Missing Data 9

2.4 Treatments for Handling Missing Data 13

2.4.1 Filter-based Method 14

2.4.2 Imputation Method 15

2.5 An Overview of Fuzzy Theory 18

2.6 Fuzzy Clustering 20

2.6.1 Fuzzy C-Means Clustering 21

2.7 Particle Swarm Optimization 23

2.8 An Overview of Decision Tree 26

2.9 Research Gap Discussion 29

2.10 Chapter Summary 32

CHAPTER 3 RESEARCH METHODOLOGY **33**

3.1 Introduction 33

3.2 Data Collection 34

3.2.1 Cleveland Heart Disease Dataset 35

3.2.2 Framingham Heart Data Study 37

3.3 The Proposed Method: The FCMPSO Imputation 38

3.4 Data Partitioning 42

3.5 Applying Decision Tree Classifier 43

3.6 The Performance Measurement 43

3.6.1 Performance Measure for Imputation 43

3.6.2 Performance Measure for Classification 44

3.7 Chapter Summary 45

CHAPTER 4 RESULTS AND ANALYSIS **46**

4.1 Introduction 46

4.2 Experiment 1 46

4.3 The Effects of FCMPSO with Different Missing Ratios on RMSE 47

4.4 The Effects of FCMPSO with Different Missing Ratios towards Classification Algorithm 48

4.4.1 Accuracy 49

4.4.2 Precision 52

4.5 Performance Evaluation Analysis 53

4.5.1 Accuracy 53

4.5.2 Precision 54

4.6 Summary 55

CHAPTER 5 CONCLUSION AND FUTURE WORKS 56

5.1 Introduction 56

5.2 Contribution of the Study 57

5.3 Recommendation for Future Works 58

5.4 Concluding Remarks 58

REFERENCE 59

APPENDIX 69

VITA 70

LIST OF TABLES

2.1	Example of training set	10
2.2	Example of testing set	10
2.3	Example of missing attributes in testing data	11
2.4	Example of Listwise deletion	15
2.5	Example of Pairwise deletion	15
2.6	Example of Mean Imputation	16
2.7	The summary of Imputation Techniques	31
3.1	Summary of Heart Disease dataset	35
3.2	Missing Percentage for dataset	36
3.3	Framingham Heart Disease dataset	37
3.4	Sample of Heart Disease dataset	41
3.5	Sample of complete data from Heart Disease dataset	41
3.6	Sample of Incomplete Data from Heart Disease dataset	41
3.7	Centroid of two clusters	42
3.8	Membership Function of complete data from Heart Disease dataset	42
4.1	Accuracy between 0% and 30% Missing Ratios	47
4.2	RMSE results for imputation methods	47
4.3	Summary of Decision Tree Parameters in WEKA	49
4.4	Comparison of Decision Tree Accuracy under different missing ratios and imputation methods.	50
4.5	Precision for Decision Tree	52

LIST OF FIGURES

2.1	Stages involved in the KDD Process by Fayyad et al. (1996)	8
2.2	Categories of Missing Data (Rubin, 1976)	12
2.3	Treatments of Missing Data	14
2.4	Example on Traditional Logic	19
2.5	Example on Fuzzy Logic	19
2.6	Decision Tree for Play Tennis	27
2.7	Confusion Matrix	44
3.1	Research Methodology	34
3.2	FCMPSO Design	38
4.1	Graph comparison of FCMPSO with Mean Imputation, k-NN and FCM on Heart Disease dataset	48
4.2	Accuracy graph for Decision Tree algorithm on Heart Disease dataset for 1% until 30% missing ratios	50
4.3	Precision graph for Decision Tree algorithm on Heart Disease dataset for 1% until 30% missing ratios	52
4.4	Decision Tree’s Accuracy	53
4.5	Decision Tree’s Precision	55

LIST OF ALGORITHMS

2.1 FCM Algorithm 21

2.2 PSO Algorithm 25

3.1 FCMPSO Algorithm 39

LIST OF SYMBOLS AND ABBREVIATIONS

WHO	-	World Health Organization
FCM	-	Fuzzy C-Means
PSO	-	Particle Swarm Optimization
FCMPSO	-	Fuzzy C-Means Particle Swarm Optimization
RMSE	-	Root Mean Square Error
UCIMLR	-	University California Irvine Machine Learning Repository
WEKA	-	Waikato Environment Knowledge Analysis
KDD	-	Knowledge Discovery in Database
MCAR	-	Missing completely at random
MAR	-	Missing at random
NMAR	-	Not missing at random
k-NN	-	*k*-Nearest Neighbor
WDS	-	World Data Strategy
PDS	-	Partial Distance Strategy
OCS	-	Optimal Completion Strategy
NPS	-	Nearest Prototype Strategy
SI	-	Swarm Intelligence
ACO	-	Ant Colony Optimization
BCO	-	Bee Colony Optimization
SVM	-	Support Vector Regression
GA	-	Genetic Algorithm
MSE	-	Mean Squared Error
AAELM	-	Auto Associative Extreme Machine Learning
EM	-	Expectation Maximation

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A.1	Gantt Chart of Research Activities	69

LIST OF PUBLICATIONS

Conference:

(i) Nurul Ashikin Samat, Mohd Najib Mohd Salleh. (2016). “Improve Decision Tree Classifier with FCMPSO for Detecting Heart Disease.” Second International Conference on Soft Computing and Data Mining (SCDM-2016)

(ii) Mohd Najib Mohd Salleh, Nurul Ashikin Samat. (2017). “FCMPSO: An Imputation for Missing Data Features in Heart Disease Classification.” International Research and Innovation Summit (IRIS17)

(iii) Nurul Ashikin Samat, Mohd Najib Mohd Salleh. (2017). “An Imputation for Missing Data Features based on Fuzzy Swarm Approach in Heart Disease Classification.” The Eighth International Conference on Swarm Intelligence (ICSI 2017)

LIST OF AWARD

(i) **2nd Place in Three Minute Thesis Competition [3MT 2016]:**

Nurul Ashikin Samat, Assoc. Prof Dr. Mohd Najib Mohd Salleh. “Improve Decision Tree Classifier with FCMPSO for Detecting Heart Disease.”

CHAPTER 1

INTRODUCTION

1.1 Introduction

In recent decades, information technology areas have been thriving worldwide. In addition, fast development of powerful data collection and storage tools contribute to the growth of available data volume. These data come from various areas and industries such as business, engineering, telecommunication, medical and health industry. Business industry generates data sets from sales transaction, stock trading, performances and customer feedbacks. While, in medical and health industry data was generated from medical records, patient monitoring and medical imaging. These databased were collected but raw data do not give any specific and important knowledge to experts. Therefore, strong and powerful tools are needed to extract and uncover the valuable knowledge from the data. This condition has demanded to the needs of data mining.

Data mining is a tool that has ability for turning the huge data into useful knowledge and information. Therefore, it has gained a lot of attention from various industries and areas in recent years. Knowledge and information extracted from the data mining gives benefits to respected industries as it can provide the sufficient evidence, indication and support for an organization to make any decisions further. There are variety of data mining tools such as classification, clustering, regression and association. Moreover, data mining implementation possesses capabilities to facilitate quality support, improved data management, and enhanced communication and production field.

Data mining classification has been widely used by researchers in various areas because it classifies given attribute for certain classes and translates the knowledge in the rules form. There are several famous techniques such as Naïve Bayesian Classifier (Muhammed, 2012), Decision Tree (Mahmood & Kuppa, 2010) (Srinivas *et al.*, 2010), Support Vector Machine (Soman *et al.*, 2003), and Neural Network (Khemphila & Boonjing, 2011) have been used before in their studies. Classification design consists of two phases: (1) Training and (2) Testing (Yoo *et al.*, 2012). However, each technique gives different accuracy respectively.

In general, Decision Tree classifier becomes a popular and competent classification technique among the researchers (Mohamed *et al.*, 2012; Tomar & Agarwal, 2013; Tsang *et al.*, 2011). Decision Tree has been applied to predict new data into the respectively class presented in tree structure. Decision Tree algorithms such as ID3, C4.5, and C5.0 have also been used widely. The algorithms use divide and conquer technique which starts with the root and moves through the branch until the node is reached. Basically, Decision Tree is practical, easy to implement and the rules extraction are easy to understand. However, the Decision Tree tends to grow a large tree with complex rules of extraction. Thus, pruning method is needed to overcome this drawback. Other than that, Decision Tree also needs certain data for classification as uncertain data can affect the accuracy of Decision Tree.

Nevertheless, the real-world data stored in a database, generally may contain noise, incomplete data, and inconsistent. These conditions may confuse the data mining process, causing the knowledge extracted unreliable. Thus, the accuracy of uncovered knowledge can be poor. As an example, in health industry, uncertain data can appear from the data collection process such as irrelevant input features, no value or missing values of input and impossible or unlikely values of input. These problems can cause the accuracy of data mining as it cannot perform well due to the incomplete features. Although data mining algorithm such as Decision Tree has its own mechanism to handle missing value as probabilistic approach, but it still does not give the best treatment to the missing data (Song *et al.*, 2008).

Thus, preprocessing method is a necessary step to address these problems. Preprocessing stage has been applied as it is an important step to improve the ability of the data mining tools to perform better and to maximize the extraction knowledge from the data itself (Tanasa & Trousse, 2004). Utilizing imputation method for missing data problem is common used among the researchers. Imputation is a process when the missing value is replaced with new value.

1.2 Problem Statement

Decision Tree has been frequently used in healthcare, manufacturing, and business to help decision maker in making an effective decision (Tsang *et al.*, 2011). The clear visualization of tree gives advantages to the user to identify the most important class. However, Decision Tree works with precise and known data to give better results (Sutton-Charani *et al.*, 2013). Thus, there are problems in decision making when there are imprecise data. Imprecise data need to be taken seriously as these data can affect the quality of decisions. Imprecise data include noise or uncertain data from no value input features and missing values recorded (Kotsiantis *et al.*, 2007). It may exist from the data collection process.

Thus, to overcome this problem, preprocessing stage is considered crucial before training the data into the classifier. Choosing the right and suitable preprocessing method can improve the dataset quality. By using poor quality data, it leads to poor quality information and knowledge. For example, given missing data in customer relation service system, customers may receive many calls due to wrong grouping, plus leading to missed sales opportunity and unhappy customers.

Hence, the preprocessing stage can eventually maximize the accuracy and efficiency of machine learning techniques. The decisions made from the data is reliable and trustworthy. Therefore, most of researchers introduced imputation methods to overcome the missing dataset problems in a preprocessing stage. Imputation shows a good and competent technique in preprocessing stage. In the meantime, fuzzy concept and fuzzy theory have many advantages in dealing with data containing uncertainty, therefore fuzzy approaches have been taken into consideration to find the imputation values.

Imputation is a method that replaces or substitutes the missing value with a new value. There are existing imputation methods such as Mean, Mode and imputation based on range idea such as *k*-Nearest Neighbor (*k*-NN) and FCM imputation. Although the imputation method has been used to handle the missing data, the accuracy for the replaced value can still be improved. Recently, clustering algorithm, Fuzzy C-Means (FCM) (Bezdek *et al.*, 1984) idea demonstrated a good response in order to fill the missing data input. Although the ability of FCM to find the plausible values to impute based on the membership values makes the algorithm a reliable way of imputation, but some features may be neglected or not properly cluster, reduce the imputation accuracy or give false imputation values.

Thus, to optimize the problems, a research has been done to improve the imputation results by applying an optimization algorithm, Particle Swarm Optimization (PSO) to optimize the imputation values. PSO is mainly based on mathematical foundation and application research to prove its convergence and robustness. It had no overlapping and mutation calculation. PSO also adopts real number and gives solution directly. It was chosen due to the simple algorithm, practical to implement and give promising results. The benefits of PSO implementation is to enhance the candidates for imputation and to choose the best suits value for replacement. Apart from that, this research proves that after FCMPSO has been applied in preprocessing stage, it leads to better imputation accuracy and significantly improve the accuracy of classification algorithm.

1.3 Aim of Study

The aim of this study is to improve on the accuracy of Decision Tree classification results between incomplete and complete dataset. Therefore, this study focused on imputation method using FCM in the preprocessing stage by optimally selecting the impute data using PSO.

1.4 Objectives of the Study

In order to achieve the research aim, three research objectives are set as follows.

(i) To propose an improved imputation technique based on Fuzzy C-Means and Particle Swarm Optimization (FCMPSO).

(ii) To apply (i) for missing dataset problem in preprocessing stage to get complete dataset.

(iii) To evaluate the performance of (i) with mean imputation, *k*-NN imputation, and FCM, respectively, based on RMSE and Decision Tree accuracy.

1.5 Scope of Study

This research focuses on the improvement of the imputation method using FCM and PSO in preprocessing stage called FCMPSO. The performance of proposed method will be compared with mean imputation, *k*-NN imputation, and FCM on the Root Mean Square Error (RMSE). In addition, this research also focuses on improvement of classification results by applying Decision Tree algorithm with the complete dataset. The experiment has been trained with Decision Tree algorithm in Waikato Environment Knowledge Analysis (WEKA) version 3.6.11. The performance of classification is measured in terms of accuracy and precision.

Heart Disease dataset from University California Irvine Machine Learning Repository (UCIMLR) (Frank & Asuncion, 2010) and Framingham Heart dataset from National Institutes of Health (NIH) (Framingham Heart Study, 2016) has been chosen as samples for the training process.

1.6 Significance of Study

In order, to understand the importance of the preprocessing towards machine learning techniques, this study investigates on the effects of imputation method in preprocessing stage which focuses on FCMPSO imputation towards Decision Tree. The findings of this study will demonstrate the vital needs for data mining to have a complete dataset to get accurate knowledge. Therefore, after the preprocessing stage is carried out, the dataset will be trained on Decision Tree and the classification rules will be extracted leads to help expert to make decisions. Thus, it will enable to produce more accurate and comprehensible decisions for organization to use.

1.7 Thesis Outline

Currently, with the rapid growth of data in business, engineering, and healthcare, data mining will reveal the pattern and knowledge from the data collected. There are many classification applications and model that are employed by the experts and industries. However, there are limitations such as uncertainty, accuracy, and complexity for some models. Thus, preprocessing stage is essential in order to preserve the ability of machine learning techniques. For that reason, a study on the improvement towards imputation methods is proposed. This study works with FCMPSO methods to impute better values towards missing problems, which in turn increases the accuracy of Decision Tree algorithm.

This thesis consists of five chapters, including this Introduction chapter. The remaining part of this thesis is segmented into following order: ***Chapter 2***: Literature Review. This chapter includes an overview of data mining classification in the healthcare industry. In this chapter, concept of missing data and the imputation methods are reviewed. Furthermore, the optimization algorithm, PSO will also be reviewed in this chapter. Then, this chapter introduces a new method in improving the imputation method by proposing an algorithm. ***Chapter 3:*** Research Methodology. This chapter discusses the steps used to systematically put the study into action. Design, formulation, and implementation of dataset to optimize imputation are discussed in detail. ***Chapter 4:*** Results and Discussion. The evaluation of optimized imputation method and Decision Tree was developed in Chapter 3. The performances of the proposed method were tested for comparison. ***Chapter 5:*** Conclusions and Future Works. This chapter concludes the works done and the recommendations are described for further continuation of work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Decision Tree is widely known due to its capabilities to classify and produce rules from the dataset. The rules that have been produced are easy and practical to be used by human experts. Nevertheless, to produce robust and reliable trees for new records prediction, Decision Tree needs a complete dataset. Hence, the existence of missing data in the dataset is somewhat unavoidable. Missing data are unfavorable to researchers and experts because it may lead errors and confusion in interpreting the data. Therefore, dealing with missing data is an important issue in data mining. The literature review regarding type of missing data and type of imputation methods used to substitute the missing value is discussed. This study focuses on the imputation of missing data in preprocessing stage by clustering the features selected based on Fuzzy C-Means clustering method. Despite the ability of FCM to find the imputation value, there is weakness that can be improved in order to find the most accurate value for imputation. An overview of Decision Tree is also discussed as it has been used to validate the performance of imputation.

This chapter is organized in the following order: Section 2.2 provides an overview of data mining and Section 2.3 presents the concept of missing data. Section 2.4 discusses the treatments for addressing the missing data and the basic introduction towards fuzzy theory is elaborated in Section 2.5. Section 2.6 focuses on clustering and fuzzy idea for imputation. In Section 2.7, the fundamentals of Particle Swarm Optimization work are presented. The classification algorithm, Decision Tree will be discussed in Section 2.8. In Section 2.9, the previous proposed solution that has been done by other researchers in regards to imputation using FCM and PSO were highlighted. At the end of this chapter, the summary regarding overall literature review is made.

2.2 An Overview of Data Mining

Over the years, information technology areas have been thriving worldwide. Data collection comes from various kinds of databases. These databases were collected from various industries such as automotive and healthcare industry. The raw data do not give any specific and important knowledge to experts, thus, data mining helps to extract the information from the data.

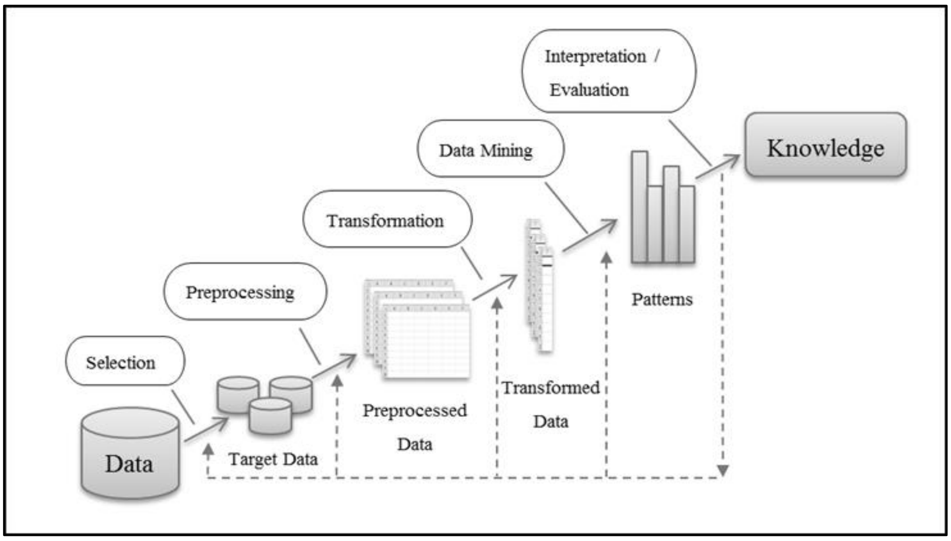

Figure 2.1: Stages involved in the KDD Process by Fayyad *et al.* (1996)

According to Fayyad *et al.* (1996), the Knowledge Discovery in Databases (KDD) needs data mining as it is an important stage for KDD to perform well. Figure 2.1 shows the five stages involved in KDD which include (1) Selection, (2) Preprocessing, (3) Transformation, (4) Data Mining, and (5) Evaluation.

There have been notable successes in the use of data mining techniques to discover scientific knowledge in the field of business, engineering and health. For an example, healthcare industry has successfully utilized the data mining method to process and analyze the huge data produced in this industry. This includes various stages in healthcare industry such as organization, management, and patients' treatments (Koh & Tan, 2011). The incorporation of computational intelligence in health diagnosis is not a new tendency. Researchers are exploiting the medical history records of the patients, so that, early detection can be performed. Therefore, it will reduce the medical costs, lessens the number of medical tests, increases the rate of successful treatments and lessen the mortality. Although data mining techniques have achieved great success, they also encountered difficulties in meeting challenges posed by the datasets. For an example, healthcare industry has unpredictable medical data where it is hard to collect accurate, precise, and complete medical data (Bratu *et al.*, 2008; Tomar & Agarwal, 2013; Wang *et al.*, 2016).

Therefore, after the selection process of desired target data in the first stage of KDD, preprocessing is an essential and important step. The dataset is filtered and cleaned before it can be trained in the data mining stage (Sridevi *et al.*, 2011). Therefore, failure to preprocess the data might affect the accuracy of data mining algorithms and affect the results of data mining analysis. This is consistent with the study by Fayyad *et al.* (1996), which mentioned that it is necessary to preprocess the data because any low quality data source such as missing value may give less optimum results to the analysis. Thus, in this research, to address missing data problems, imputation method in preprocessing stage has been selected. Meanwhile, missing data definition will be further explained in next subtopic 2.3.

2.3 Concept of Missing Data

In the real world, missing data are unfavorable by the researchers and experts because it may lead to bias, errors and confusion in interpreting the data. However, ignoring the missing data also leads to a disadvantage as it may contain other important information. Thus, it has attracted a significant research interest in recent years. In previous subchapter, it is mentioned that preprocessing stage is a way to clean and filter the target data because the data collection is not always complete and accurate (Salleh, 2013). The collection process might involve or be tangled with uncertain environments and consequently, imprecise datasets appeared. In the past years, researches have studied the effect of missing values towards the analysis process. Three types of problems are usually associated with missing values: 1) loss of efficiency; 2) complications in handling and analyzing the data; and 3) bias resulting from differences between missing and complete data. Thus, it is important to have complete and quality dataset.

In this study, classification using Decision Tree algorithm has been chosen to mine the knowledge from the data. Classification is used to classify the specified data into class label and predict the future data in the previously classified template class. Table 2.1, Table 2.2 and Table 2.3 will help to clarify briefly about classification process.

Table 2.1: Example of training set

Age	**Gender**	**Chest Pain Type**	**Blood Pressure**	**Heart Disease**
40	Male	Atypical angina	140	No
49	Female	Non- angina pain	160	Yes
37	Male	Asymptomatic	150	Yes

Table 2.1 shows the example of training set for Heart Disease patients. The table contains a set of attributes which includes age, gender, chest pain type, and blood pressure. Meanwhile, the class or goal of attribute in the training set is the presence of Heart Disease. In the first step of the classification process, the attributes in the training set are analyzed by classification algorithm and the relationship knowledge between the attributes and class is identified. The classifier model is then shown in rule or pattern form. The rule generated might be expressed like this;

IF (age > 35 & blood pressure < 150)

THEN (Heart Disease = No).

IF (chest pain type = asymptomatic & blood pressure > 150),

THEN (Heart Disease = Yes).

Table 2.2: Example of testing set

Age	**Gender**	**Chest Pain Type**	**Blood Pressure**	**Heart Disease**
48	Male	Asymptomatic	138	?
54	Female	Non- angina pain	150	?
39	Male	Non- angina pain	120	?

Table 2.2 shows the testing dataset for new Heart Disease patients with the same attributes like training data, but no class or Heart Disease is determined. As a result, the second process of classification is applied. Hence, the testing data that have been collected will be classified accordingly by applying a classification algorithm for the class or goal attribute. It is also used to estimate the classification

accuracy based on the percentages that classified correctly by the algorithm. However, what will happen if the testing data have missing attributes like Table 2.3.

Table 2.3: Example of missing attributes in testing data

Age	**Gender**	**Chest Pain Type**	**Blood Pressure**	**Heart Disease**
43	-	Asymptomatic	-	?
54	Female	-	150	?
-	Male	Non-angina pain	145	?

Table 2.3 shows an example of several missing attribute values or incomplete data in the database. These uncertain data are inherited from real-life data. Missing values in training or testing data can be disadvantageous to knowledge extraction process. Uncertain environments might arise from several factors such as parallax error, human error or equipment error (Zhang *et al.*, 2010). As an example, healthcare industry always has unpredictable medical data. It is hard to collect accurate, precise, and complete medical data (Tomar & Agarwal, 2013). Medical data records come from various materials such as medical reports, laboratory report, X-ray report, and report reviewed by experts from each appointment with patients. Although there are some ways to avoid from getting missing data such as repeating the experiment and lab test, it is not always possible to avoid missing data and this will increase the medical expenses to the patients.

There are several ways to deal with missing values in data sets. Deleting or ignoring the missing data are the simplest approaches. Nevertheless, to ignore the whole row of data which contains missing data might be an ineffective move, as other complete data can give the information in certain features or groups. For an example, in Table 2.3, for row number one, the "*Blood Pressure*" and "*Gender*" data are missing, but that row still has the complete data on "*Age*" and "*Chest Pain Type*". Thus, for features "*Age*" and "*Chest Pain Type*", they still need to be counted for analysis, as they may contain important information such as, "In which certain age range that has a high risk of Heart Disease problem".

Therefore, the missing data need to be treated with imputations. The accuracy of the imputation or substitute values are also needs to be counted, so that it can give better results and useful information. Thus, with the idea that other complete data in the missing data rows are useful, missing data can be treated by clustering the complete features to find the possible values to be imputed.

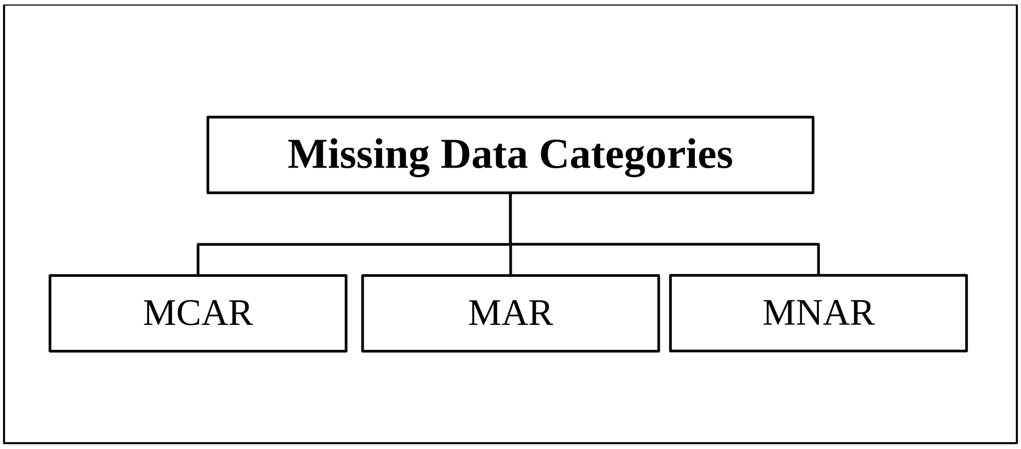

Figure 2.2: Categories of Missing Data (Rubin, 1976)

Figure 2.2 shows that according to Rubin (1976) missing data can be categorized into three which are Missing completely at random (MCAR), Missing at random (MAR), and Not missing at random (NMAR) (Rubin, 1976). Missing data is a situation in which some components of the dataset are not available for all feature variables, or may not even be defined within the problem domain. The following parts will explain briefly about each missing mechanism.

(i) Missing completely at random (MCAR)

Missing completely at random happens in a situation where the missing data in the dataset for an attribute has no relationship or dependency towards another set of attributes. An example of MCAR situation is where the expert needs to record the patient’s latest blood pressure, but he did not show up for the follow-up medical check-up due to family problems or being involved in an accident. Thus, the missing data for “blood pressure readings” attribute is absolutely not related to other attributes such as gender and age.

(ii) Missing at random (MAR)

Missing at random is the missing data in the dataset for an attribute that depends on other attributes or variables within the dataset. The probability of the missing values can be obtained by estimating other complete attributes or variables. MAR is the most general condition considered by researcher to perform an analysis of the missing value (Aydilek & Arslan, 2013).

(iii) Missing not at random (MNAR)

The third category, missing not at random is the missing value that does not fulfil the other two mechanisms. MNAR is the missing value that can be influenced and depends on the same attribute only (Leke *et al.*, 2015). It makes the missing value not random and be estimated from other attributes within the dataset.

Missing problems is a common problem in real world industries and areas such as Semiconductor industry (Azarkhail & Woytowitz, 2013), patients medical records (Rahman *et al.*, 2014), Data preprocessing (Zhang *et al.*, 2006), and Microarray experiments (Bose *et al.*, 2012). The above literature is consistent with this study where missing data may have problems in the extraction of information. Nevertheless, with the precise ways to treat the problem, it can be an advantage. Thus, in the next subchapter, the treatment methods available for treating the missing data will be further discussed and explained.

2.4 Treatments for Handling Missing Data

One of the most significant current discussions in missing data is the treatments available for it. In this section, ways to handle the missing data is discussed briefly. In the past years, researchers have been studying the missing data problems and dozens of techniques to address the problem have been reported. Many researchers and experts have agreed that poor quality data can result to less accurate and not reliable knowledge presented by the data mining model. In addition, the decision-making process by the experts are frequently getting more complex due to the missing problems. Later in this section, few common treatments of missing data are highlighted. Each method in this section deals with missing data by (1) removing the incomplete data or (2) by filling in the missing values.

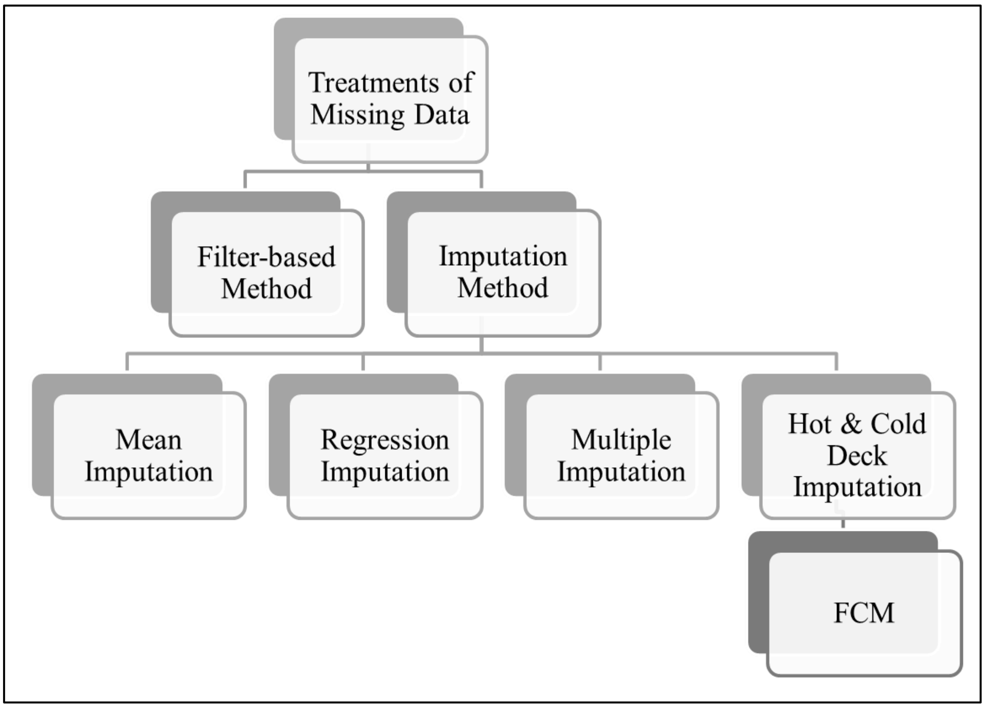

Figure 2.3: Treatments of Missing Data

There are several methods applied by the researchers as shown in Figure 2.3 to treat the missing data in the preprocessing stage. This is important as the best technique will maximize the knowledge extracted from the dataset and improved the performance of data mining algorithm. The following subsection will discuss briefly about each treatment.

2.4.1 Filter-based Method

Filter-based treatment means that the missing value is ignored or deleted. It means that no imputation value substitutes to the missing data. Before this, Rubin (1976) introduced that general way to treat the missing data which is to ignore the missing data or using deletion methods such as Listwise deletion and Pairwise deletion (Karadog *et al.*, 2011).

Listwise deletion is the simplest way of deleting records that contain missing values. Subsequently, the number of records will be less and the information cannot be fully utilized like Table 2.4.

Table 2.4: Example of Listwise deletion

ID	Attribute 1	Attribute 2
A1	23	154
~~A2~~	~~34~~	~~?~~
A3	60	89
A4	45	112
~~A5~~	~~?~~	~~164~~

Table 2.5: Example of Pairwise deletion

ID	Attribute 1	Attribute 2
A1	23	154
A2	34	~~?~~
A3	60	89
A4	45	112
A5	~~?~~	164

On the other hand, Table 2.5 shows that pairwise deletion only deletes the missing data without deleting the whole record. Nevertheless, the analysis part will have problems as the number of each sample is different in the dataset.

Thus, both ways are acceptable and suitable if the missing data is in a small number. Even though this method is more stable (Karadog *et al.*, 2011), a bigger number will affect the quality of the dataset and it may result to biased data (Aydilek & Arslan, 2013; Dhevi, 2014; Thirukumaran & Sumathi, 2012). Apart from that, deletion methods will also give lower precision values over the increasing missing values in the dataset (Twala *et al.*, 2005). Furthermore, the real world database will has a relatively a large number of data, and thus filter-based method is not a practical method to implement (Farhangfar *et al.*, 2007).

2.4.2 Imputation Method

An imputation method is where the missing values are imputed with estimated values by using the information from the complete dataset or complete instances. There are many imputation methods that had been proposed by researches. Authors Sridevi *et al.* (2011) have implement imputation method in time series data based on history data. Authors believed that temporal data mining that relates to time information is also important in implementing imputation method. As for Chien-Lung *et al.* (2011), authors mentioned that in getting good and reliable information for traffic management system, missing traffic data can lead to wrong conclusion or information. Thus, by implementing imputation method, the information extraction can be carried out better. As for medical field, Zhang *et al.* (2012) applied imputation method to help the classifier to overcome the missing problem in clinical heart failure data.

Although, researchers give good reviews about imputation in each their case study in making the data mining work better. However, getting the best imputation values to replace the missing data is still very challenging. The following subsection will discuss each imputation method briefly.

(i) Mean imputation

First imputation is Mean imputation which is the earliest method of imputation (Ravi & Krishna, 2014). The missing value is replaced with the mean value for the attribute or variables. Although it is easy to use, it will affect the relationship between attributes or variables, and weakens the covariance and correlation in estimation as each missing value is imputed with the same imputed value (Enders, 2010). Example of Mean imputation is as follow;

Table 2.6: Example of Mean Imputation

ID	Attribute 1	Attribute 2
A1	23	154
A2	34	?
A3	60	89
A4	45	112
A5	?	164

From Table 2.6, ID A2 and A5 have a missing value in Attribute 2 and Attribute 1 respectively. Thus, for Mean imputation, the value is replaced by 129.75 for A2 and 40.5 for A5. Below is the calculation for Mean imputation.

Mean value for ID A2;Attribute 2 = (154 + 89 + 112 + 164) ÷ 4 = 129.75	Mean value for ID A5;Attribute 1 = (23 + 34 + 60 + 45) ÷ 4 = 40.5

(ii) Regression imputation

Second imputation is the regression imputation where the missing data is imputed using predicted score from the regression equation. Regression imputation utilizes the idea to find the imputation value from the results of regression complete values as they might have relation between the attributes (Enders, 2010). Apart from that, regression model has also been used together with prediction model such as Neural Network to improve the estimation data (Lingras *et al.*, 2008). Although regression can impute better than mean imputation, the variability of the imputation range is too small. Thus, it can lead to wrong inferences from the dataset. Apart from that, regression is also tangled with biased problem estimation (Shao & Wang, 2002).

(iii) Multiple imputation

Third imputation is multiple imputation where the basic idea of this imputation is to generate a small copy of the dataset, n (example: 5-10 subsets) which contains the missing data. The estimate value for missing data in each small copy is imputed into the missing data and result in complete full data. This activity is performed multiple times according to the *n* values. Each imputed dataset will be analyzed accordingly and then, all the results are combined to produce overall analysis (Royston, 2004). Maximum likelihood algorithm such as Expectation maximization algorithm is always used in multiple imputation (Enders, 2001). However, this type of imputation needs a lot of costs and is not easy to implement as they are embedded into costly software (Myers, 2011).

(iv) Hot deck and cold deck imputation

The idea of hot deck and cold deck imputation is to impute the missing data with an actual range of datasets (Roth, 1994). The difference between them is that hot deck uses information to impute from the same dataset, while cold deck uses other datasets to impute. According to Myers (2011), hot deck imputation is easy and less costly to implement in various areas of study and environment. Apart from that, hot deck imputation does not give the out range of imputed values such as multiple imputations and it does not need to define any model for the placement of missing data. The *k*-Nearest Neighbor (*k*-NN) imputation method is a common hot deck method (Jönsson & Wohlin, 2004). The *k*-NN uses the complete dataset to find the plausible neighbors to impute into the missing data based on the distance between them. However, this method takes more computational time for larger dataset as *k*-NN looks for the most similar instances, thus the algorithm searches through all the datasets (Batista & Monard, 2002). For hot deck imputation, to get a better range of dataset values to substitute, grouping or clustering the data with same similarity features or data will increase the accuracy of imputation values. Two examples of hot deck imputation based on clustering methods are K-Means clustering and Fuzzy C-Means clustering (Aydilek & Arslan, 2013).

In the past years, researchers had use and utilize either the regression idea or the clustering idea to identify the best imputation values by discovering the suitable range for replacement. Therefore, the focus of this study is to handle the missing data using a hot deck imputation method through clustering method. However, the missing data contained uncertainty and imprecise information. Thus, fuzzy capabilities are introduced in the next subchapter to solve the problems.

2.5 An Overview of Fuzzy Theory

Over the last few decades, fuzzy logic has been shown as a great approach for dealing with imprecision and nonlinearity efficiently. Applications can be found in a wide perspective ranging from medication to economics, supply chain management to user products, and air conditioning control to traffic control (Akhoondi & Hosseini, 2016; Al-Awadhi *et al.*, 2015; Ayağ *et al.*, 2013; Jianzhong & Jundan, 2015; Kaur & Kaur, 2012; Zhang *et al.*, 2013).

The concept of fuzzy logic was introduced by Zadeh (1965). Fuzzy logic is an approach to computing based on “degree of truth” rather than the usual “true or false” (1 or 0). The concept of information is inherently associated with the concept of uncertainty. The most fundamental aspect is that the uncertainty involved in any problem-solving situation is a result of some information deficiencies, which may be incomplete, imprecise, fragmentary, not fully reliable, vague, contradictory, or deficient in some other ways. Generally, fuzzy allows handling much of this uncertainty, which represents uncertainty by numbers in the range between 0 until 1. Figure 2.4 and Figure 2.5 illustrates a simple example of the traditional and fuzzy logic towards attribute blood pressure.

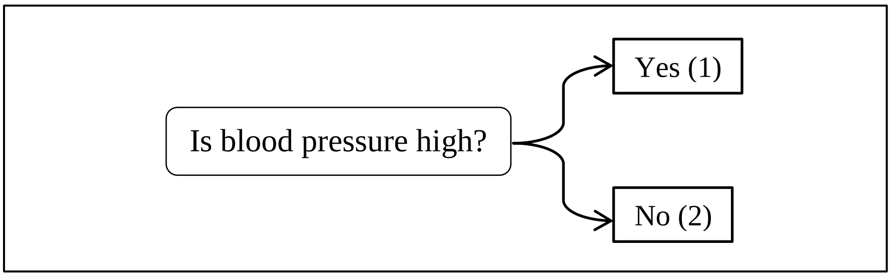

Figure 2.4: Example on Traditional Logic

Figure 2.4 shows traditional or binary logic answer towards blood pressure where there are only two answers which are "Yes" and "No". Therefore, the binary answer is referred to as "Yes" to represent the value of "1" and "No" to represent the value of "0". Most traditional tools for modelling, reasoning and computing are crisp, deterministic, and precise in character. In conventional dual logic for instance, a statement can be true or false, and nothing in between.

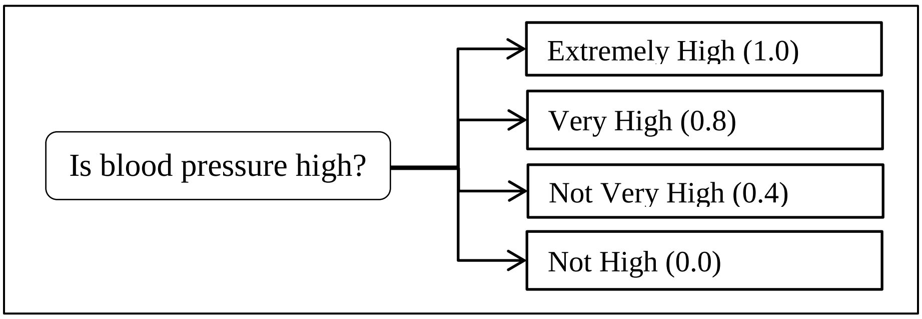

Figure 2.5: Example on Fuzzy Logic

On the other hand, Figure 2.5 shows multiple answers towards blood pressure question which implement the concept of fuzzy logic. "Extremely High" represents the value of 1.0, "Very High" represents the value of 0.8, "Not Very High" represents the value of 0.4 and "Not High" represents the value of 0.0. This is degree of truth and the value ranges from 0 to 1. It is called as membership function values. Since its appearance, the theory of fuzzy has advanced in a variety of ways and in many disciplines. This trend is still ongoing as fuzzy is the best tools for modelling uncertain problem solving. In this study, fuzzy theory is used to accommodate fuzziness in human judgement, evaluation and decision. It is important to develop better understanding for machine intelligence to mimic human decision-making. Thus, non-expert also can use the knowledge to make decisions.

The classical methods for data mining, such as clustering techniques, are available, but sometimes they do not match the needs. For instances, although clustering techniques, assume that data could be subdivided crisply into clusters, they would not fit the structures that exist in reality. Fuzzy set theory seems to offer good opportunities to improve existing concepts.

2.6 Fuzzy Clustering

Clustering is one of the methods in data mining. The purpose of clustering is to group the dataset into subsets based on their similarity. Since the imputation value is substituted with the actual range of the feature respectively, clustering is a suitable approach to find the group of useful features by having the similarity measure. In order to find the plausible value for missing data, clustering uses a local modelling technique which is grouping the data and attributes that have connection with the missing data first and then, the imputed value is calculated from the subset group. It is applicable to real life problem as well, for example, Heart Disease patient data in the hospital database. The database can contain attributes such as age, gender, blood pressure, chest pain type and etc. Based on these attributes, it is possible to find the connection of patients into several clusters where they have similar features.

With the clustering knowledge, the imputation of missing values can be done with higher accuracy by selecting the meaningful, useful and significant features. Clustering has a predictive power, allowing one to predict those data patterns that share the same cluster to also have similar properties. A summary of clusters can be communicated using cluster centers, which are representatives of the clusters. It is important to note that, clustering does not have its own mechanism to handle the missing value. However, it is a common solution used by other researchers to help the imputation in the preprocessing stage because of the ability of clustering algorithm to group the complete features and find the connection between the features subsets. In addition, the process of cluster formation is unaffected of information about the data sources such as class labels, which influence the interpretation of results later. For the past year, researchers have come out with imputation ideas and implementation in various areas of studies. Thus in the next subchapter, fuzzy clustering algorithms, Fuzzy C-Means will be explained further.

2.6.1 Fuzzy C-Means Clustering

The best known and the most widely used fuzzy clustering algorithm is the Fuzzy C-Means clustering (FCM) algorithm. In this study, FCM was chosen as clustering algorithm to find the new value for missing data. FCM has a high reputation as a clustering algorithm and has various implementations and variations. FCM was developed by Dunn in 1972 and improved by Bezdek in 1981 (Cannon *et al.*, 1986). The fuzzy clustering algorithm allows one data to belong to two or more clusters (Jiawei *et al.*, 2005). The standard clustering algorithm partitions data in "either-or" type of division, in which the membership function of the data is either 0 or 1 only. Thus FCM extends the membership function for the data so that it can have a value from 0 to 1 which can improve the clustering results (Niu & Huang, 2011).

FCM is different in terms of its implementation of fuzzy model. It does not include a fuzzifier, fuzzy rules, fuzzy inference engine, and defuzzifier steps, and does not produce IF – THEN rules. It is an iterative algorithm that minimizes the objective function to cluster data with better grouping. The algorithm is iterative and can be stated as follows:

```
Start
    Step 1: Set the number of cluster and fuzzy index $m$ $(m > 1)$, initializes
            the matrix of membership function values $\mu_{ij}$, set maximum
            iterations $w$.
    Step 2: Compute the cluster centres, $r_j, j = 1, 2, \dots, c$ according to
            Equation 2.5
    Step 3: Compute Euclidean distance and Update the membership function
            $\mu_{ij}$ according to:                                    (2.1)
            $\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}}$
    Step 4: If not converged, repeat step 2 until the completion of maximum
            number of iterations.
End
```

Algorithm 2.1: FCM Algorithm

FCM partitions set of n dataset $x = \{x_1, x_2, \ldots, x_n\}$ in R^d dimensional space into fuzzy cluster c, $1 < c < n$ with $r = \{r_1, r_2, \ldots, r_c\}$ cluster centers or centroids. The fuzzy clustering dataset is described by fuzzy matrix μ with n rows and c columns in which n is number of dataset and c is the number of clusters. Meanwhile, μ_{ij} is the element in i^{th} row and j^{th} column in μ, showing the membership function of the i^{th} dataset with the j^{th} cluster. μ is defined as follows,

$$\mu_{ij} \in [0,1] \; \forall i = 1,2,\ldots n; \; \forall j = 1,2,\ldots, c \tag{2.2}$$

$$\sum_{j=1}^{c} \mu_{ij} = 1, \; \forall i = 1, 2, \ldots, n \tag{2.3}$$

The objective function of FCM algorithm to minimize iteratively,

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{m} \cdot d^2(x_i, r_j) \tag{2.4}$$

In which $m (m > 1)$ is a scalar term for the weighting exponent that controls the fuzziness of the resulting clusters and $d^2(x_i, r_j)$ stands for the Euclidean distance from dataset x_i to the cluster center r_j. The centroid r_j of the j^{th} cluster is obtained using,

$$r_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \tag{2.5}$$

However, when implementing fuzzy algorithm, it is important to choose an appropriate value for parameters such as the fuzziness exponent m especially, in fuzzy models as the minimization criterion for the objective function depends on m. The m parameter determines the vagueness of the resulting partitioning. According to Berget *et al.* (2008) a value of $m = 2$ is commonly used and an increased value of m can be interpreted as an increased sharing of points among all clusters. This is also to avoid complicated computation which leads to the consumption of time. Apart from that, this value has also been proven to give good results with FCM (Berget *et al.*, 2008).

The ability of FCM has been implemented in various areas, especially in the healthcare industry and studies. Mohan and Moorthy (2013) proposed that FCM can be used to find the ratios for early detection of Diabetic Retinopathy Edema. It increases the detection of disease in the early stage and decreases the loss vision risk for patients. Ferreira *et al.* (2015) proposed that FCM is a better clustering algorithm for identifying variant and invariant medical features for Intensive Care Unit (ICU) cases. Menon and Ramakrishnan (2015) demonstrated that FCM helps to identify brain tumor using MRI Brain Image segmentation.

In 2001, Hathaway and Bezdek listed four ways to utilize the FCM algorithm for missing data problems. The simple strategy is whole data strategy (WDS) that removes all sample data that include missing values from the dataset and apply FCM to the remaining complete data, but the strategy is not desirable because the elimination leads to the loss of information. Another method that uses the partial distance strategy (PDS) calculates partial distances using all available attribute values. The third which is the optimal completion strategy (OCS) views the missing values as an optimization problem and imputes missing data in each iteration to find better estimates. The nearest prototype strategy (NPS) replaces missing values with the corresponding attributes of the nearest prototype.

Thus, in this study, the proposed approach will utilize PDS strategy in FCM algorithm and optimize the approach using Particle Swarm Optimization (PSO). Apart from that, when using FCM, even for non-fuzzy data that do not have the fuzziness and the membership value, can be set between 0 and 1 where value 1 is for the cluster and 0 is for other clusters (Rahman & Islam, 2015). Hence, next subchapter will explain more details regarding the usage of Particle Swarm Optimization to optimize the imputed value based on FCM algorithm as mentioned in Chapter 1.

2.7 Particle Swarm Optimization

For years, researchers have shown interest and turned focus to population-based algorithm or swarm intelligence. Swarm intelligence (SI), was inspired by the biological behavior of animals, and is an innovative distributed intelligent paradigm for solving optimization problems. It is a famous technique for data mining which has been found efficient for clustering and classification. Famous intelligence swarms such as Particle Swarm Optimization (PSO) (Eberhart & Kennedy, 1995), Ant Colony Optimization (ACO) (Dorigo *et al.*, 2006), Bee Colony Optimization (BCO) (Karaboga & Basturk, 2007), and Cuckoo Search (Yang & Deb, 2009) have been applied to handle various optimization problems.

The optimization is important as these algorithms help to find the best values for the problems under special and specified conditions (Civicioglu & Besdok, 2013). Due to the simplicity of the framework of PSO, the algorithms can find the optimization solution directly within acceptable computation time (Rana *et al.*, 2011; Rasip *et al.*, 2015) . Thus, with this ability, PSO has been widely implemented in many different areas.

Particle Swarm Optimization was developed and introduced by Kennedy and Eberhart in 1995 based on the natural behavior of bird flocking or fish schooling to find food (Eberhart & Kennedy, 1995). The ability of natural behavior to work as a group to find the desired point has triggered the idea to implement it for solving many problems. A flock of bird flying in a group follows the member that has the closest distance to the destination. They are travelling in a group without ever colliding with one another. Each of the members can adjust its position and velocity using the information from the group.

PSO is a heuristic search algorithm that optimizes the search space to find good solutions. In traditional PSO, population is called the swarm and the candidate of solutions in swarm is called particles. Each element of the particle contains parameter; own position, velocity, and historical information. Each particle is given random position in search space and random velocity for the particles to fly within the search space.

The particles are moved around according to few simple update algorithms. The algorithm then updates the particle in swarm by updating the velocity and position in each particle. The movements of the particles have been guided by the particle’s own best position (pbest) and the swarm best position (gbest).

The basic process of the PSO algorithm is given as follows.

REFERENCES

Abobaker, R. A., Ayob, M., & Hadwan, M. (2011). Greedy constructive heuristic and local search algorithm for solving Nurse Rostering Problems. *Proceedings of the Data Mining and Optimization (DMO), 2011 3rd Conference on*. pp. 194-198.

Akhoondi, R., & Hosseini, R. (2016). A Fuzzy Expert System for Prognosis of the Risk of Development of Heart Disease. *Journal of Advances in Computer Research, 7 (2),* pp. 101-114.

Al-Awadhi, F., Yousef, M. A., & Alkandari, A. (2015). Dynamic Fuzzy Logic Traffic Light Integrated System with Accident Detection System Using iTraffic Simulation. *Proceedings of the 2015 4th International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*. IEEE. pp. 140-145.

Amiri, M., & Jensen, R. (2016). Missing data imputation using fuzzy-rough methods. *Neurocomputing, 205* pp. 152-164.

Ayağ, Z., Samanlioglu, F., & Büyüközkan, G. (2013). A fuzzy QFD approach to determine supply chain management strategies in the dairy industry. *Journal of Intelligent Manufacturing, 24 (6),* pp. 1111-1122.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences, 233 (0),* pp. 25-35.

Azarkhail, M., & Woytowitz, P. (2013). Uncertainty management in model-based imputation for missing data. *Proceedings of the Reliability and Maintainability Symposium (RAMS), 2013 Proceedings - Annual*. pp. 1-7.

Batista, G. E., & Monard, M. C. (2002). A Study of K-Nearest Neighbour as an Imputation Method. *HIS, 87 (251-260),* pp. 48.

Berget, I., Mevik, B.-H., & Næs, T. (2008). New modifications and applications of fuzzy -means methodology. *Computational Statistics & Data Analysis, 52 (5)*, pp. 2403-2418.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10 (2)*, pp. 191-203.

Bose, S., Das, C., Dutta, S., & Chattopadhyay, S. (2012). A novel interpolation based missing value estimation method to predict missing values in microarray gene expression data. *Proceedings of the Communications, Devices and Intelligent Systems (CODIS), 2012 International Conference on*. pp. 318-321.

Bratu, C. V., Muresan, T., & Potolea, R. (2008). Improving classification performance on real data through imputation. *Proceedings of the Automation, Quality and Testing, Robotics, 2008. AQTR 2008. IEEE International Conference on*. pp. 464-469.

Cannon, R. L., Dave, J. V., & Bezdek, J. C. (1986). Efficient implementation of the fuzzy c-means clustering algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, (2)*, pp. 248-255.

Chen, L., Bi, L., Si, H., Zhang, J., & Ren, Y. (2015). Research on prediction method for pivotal indicator of hospital medical quality using decision tree. *Proceedings of the Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on*. pp. 247-250.

Chen, M. C., Liao, H. C., & Huang, C. L. (2006). Predicting Breast Tumor via Mining DNA Viruses with Decision Tree. *Proceedings of the Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*. pp. 3585-3589.

Chien-Lung, C., Cheng-Yang, L., Nan-Ping, Y., & Sheng-Yuan, S. (2011). Classification Method Incorporating Decision Tree with Particle Swarm Optimization. *Proceedings of the Genetic and Evolutionary Computing (ICGEC), 2011 Fifth International Conference on*. pp. 216-219.

Civicioglu, P., & Besdok, E. (2013). A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. *Artificial Intelligence Review, 39 (4)*, pp. 315-346.

Dass, M. V., Rasheed, M. A., & Ali, M. M. (2014). Classification of lung cancer subtypes by data mining technique. *Proceedings of the Control,* *Instrumentation, Energy and Communication (CIEC), 2014 International Conference on*. pp. 558-562.

Dhevi, A. T. S. (2014). Imputing missing values using Inverse Distance Weighted Interpolation for time series data. *Proceedings of the 2014 Sixth International Conference on Advanced Computing (ICoAC)*. pp. 255-259.

Di Nuovo, A. G. (2011). Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario. *Expert Systems with Applications, 38 (6)*, pp. 6793-6797.

Dorigo, M., Birattari, M., & Stützle, T. (2006). Ant colony optimization. *Computational Intelligence Magazine, IEEE, 1 (4)*, pp. 28-39.

Eberhart, R. C., & Kennedy, J. (1995). A new optimizer using particle swarm theory. *Proceedings of the Proceedings of the sixth international symposium on micro machine and human science*. New York, NY. pp. 39-43.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8 (1)*, pp. 128-141.

Enders, C. K. (2010). *Applied missing data analysis*. New York,NY: The Guilford Press.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 37 (5)*, pp. 692-709.

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

Ferreira, M. C., Salgado, C. M., Viegas, J. L., Sch, H., x00E, fer *et al.* (2015). Fuzzy modeling based on Mixed Fuzzy Clustering for health care applications. *Proceedings of the Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*. pp. 1-5.

Frank, A., & Asuncion, A. (2010). UCI machine learning repository. pp.

Gautam, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing, 156* pp. 134-142.

Guleria, P., Thakur, N., & Sood, M. (2014). Predicting student performance using decision tree classifiers and information gain. *Proceedings of the Parallel, Distributed and Grid Computing (PDGC), 2014 International Conference on*. pp. 126-129.

Jabbar, M. A., Deekshatulu, B. L., & Chndra, P. (2014). Alternating decision trees for early diagnosis of heart disease. *Proceedings of the Circuits, Communication, Control and Computing (I4C), 2014 International Conference on*. pp. 322-328.

Jianzhong, Y., & Jundan, P. (2015). A Research on the Financing Efficiency of Small and Medium-Sized Enterprises by Fuzzy Evaluation Method. *Proceedings of the 2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. pp. 89-93.

Jiawei, L., TaoYang, & YanWang (2005). Missing value estimation for microarray data based on fuzzy C-means clustering. *Proceedings of the High-Performance Computing in Asia-Pacific Region, 2005. Proceedings. Eighth International Conference on*. pp. 6 pp.-616.

Jönsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *Proceedings of the Software Metrics, 2004. Proceedings. 10th International Symposium on*. IEEE. pp. 108-118.

Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of global optimization, 39 (3),* pp. 459-471.

Karadog, S. G., x, an, Marchegiani, L., Hansen, L. K., & Larsen, J. (2011). How efficient is estimation with missing data? *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2260-2263.

Kaur, A., & Kaur, A. (2012). Comparison of fuzzy logic and neuro-fuzzy algorithms for air conditioning system. *International journal of soft computing and engineering, 2 (1),* pp. 417-420.

Khare, A., & Rangnekar, S. (2013). A review of particle swarm optimization and its applications in Solar Photovoltaic system. *Applied Soft Computing, 13 (5),* pp. 2997-3006.

Khemphila, A., & Boonjing, V. (2011). Heart Disease Classification Using Neural Network and Feature Selection. *Proceedings of the Systems Engineering (ICSEng), 2011 21st International Conference on*. pp. 406-409.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management, 19 (2),* pp. 65.

Krishna, M., & Ravi, V. (2013). Particle swarm optimization and covariance matrix based data imputation. *Proceedings of the Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on*. IEEE. pp. 1-6.

Leke, C., Marwala, T., & Paul, S. (2015). Proposition of a Theoretical Model for Missing Data Imputation using Deep Learning and Evolutionary Algorithms. *arXiv preprint arXiv:1512.01362,* pp.

Li, D., Gu, H., & Zhang, L. (2010). A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. *Expert Systems with Applications, 37 (10),* pp. 6942-6947.

Lingras, P., Zhong, M., & Sharma, S. (2008). Evolutionary regression and neural imputations of missing values. in (Eds.). *Soft Computing Applications in Industry*. Springer. pp. 151-163.

Liu, Y. Q., Wang, C., & Zhang, L. (2009). Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data. *Proceedings of the Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on*. pp. 1-4.

Mahmood, A. M., & Kuppa, M. R. (2010). Early Detection of Clinical Parameters in Heart Disease by Improved Decision Tree Algorithm. *Proceedings of the Information Technology for Real World Problems (VCON), 2010 Second Vaagdevi International Conference on*. pp. 24-29.

Menon, N., & Ramakrishnan, R. (2015). Brain Tumor Segmentation in MRI images using unsupervised Artificial Bee Colony algorithm and FCM clustering. *Proceedings of the Communications and Signal Processing (ICCSP), 2015 International Conference on*. pp. 0006-0009.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning, 4 (2),* pp. 227-243.

Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2012). A comparative study of Reduced Error Pruning method in decision tree algorithms. *Proceedings of the Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on*. pp. 392-397.

Mohan, A., & Moorthy, K. (2013). Early detection of diabetic retinopathy edema using FCM. *International Journal of Science and Research (IJSR), India, 2 (5),* pp. 115-118.

Muhammed, L. A. N. (2012). Using data mining technique to diagnosis heart disease. *Proceedings of the Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on*. pp. 1-3.

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures, 5 (4),* pp. 297-310.

Nishanth, K. J., Ravi, V., Ankaiah, N., & Bose, I. (2012). Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. *Expert Systems with Applications, 39 (12),* pp. 10583-10589.

Niu, Q., & Huang, X. (2011). An improved fuzzy C-means clustering algorithm based on PSO. *Journal of Software, 6 (5),* pp. 873-879.

Nugroho, F. X. S. D., Adji, T. B., & Fauziati, S. (2014). Decision support system for stock trading using multiple indicators decision tree. *Proceedings of the Information Technology, Computer and Electrical Engineering (ICITACEE), 2014 1st International Conference on*. pp. 291-296.

Poonkuzhali, S., Kumar, R. K., & Viswanathan, C. (2015). Law Reckoner for Indian Judiciary: An Android Application for Retrieving Law Information Using Data Mining Methods. in (Eds.). *Advanced Computer and Communication Engineering Technology*. Springer. pp. 585-593.

Purushottam, Saxena, K., & Sharma, R. (2015). Efficient heart disease prediction system using decision tree. *Proceedings of the Computing, Communication & Automation (ICCCA), 2015 International Conference on*. pp. 72-77.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1 (1),* pp. 81-106.

Rahman, M. G., & Islam, M. Z. (2015). Missing value imputation using a fuzzy clustering-based EM approach. *Knowledge and Information Systems,* pp. 1-34.

Rahman, S. A., Huang, Y., Claassen, J., & Kleinberg, S. (2014). Imputation of Missing Values in Time Series with Lagged Correlations. *Proceedings of the Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. pp. 753-762.

Rajendran, P., Madheswaran, M., & Naganandhini, K. (2010). An improved pre-processing technique with image mining approach for the medical image classification. *Proceedings of the Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on*. IEEE. pp. 1-7.

Rana, S., Jasola, S., & Kumar, R. (2011). A review on particle swarm optimization algorithms and their applications to data clustering. *Artificial Intelligence Review, 35 (3),* pp. 211-222.

Rasip, N. M., Basari, A., Ibrahim, N. K., & Hussin, B. (2015). Enhancement of Nurse Scheduling Steps Using Particle Swarm Optimization. in (Eds.). *Advanced Computer and Communication Engineering Technology*. Springer. pp. 459-469.

Ravi, V., & Krishna, M. (2014). A new online data imputation method based on general regression auto associative neural network. *Neurocomputing, 138* pp. 106-113.

Röhler, A., & Chen, S. (2011). An analysis of sub-swarms in multi-swarm systems. *AI 2011: Advances in Artificial Intelligence,* pp. 271-280.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel psychology, 47 (3),* pp. 537-560.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal, 4 (3),* pp. 227-241.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63 (3),* pp. 581-592.

Salleh, M. N. M. (2013). Implementing fuzzy modeling of decision support for crop planting management. *Proceedings of the Fuzzy Theory and Its Applications (iFUZZY), 2013 International Conference on*. pp. 161-166.

Shao, J., & Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association, 97 (458),* pp. 544-552.

Soman, K. P., Shyam, D. M., & Madhavdas, P. (2003). Efficient classification and analysis of ischemic heart disease using proximal support vector machines based decision trees. *Proceedings of the TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*. pp. 214-217 Vol.211.

Song, Q., Shepperd, M., Chen, X., & Liu, J. (2008). Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation. *Journal of Systems and software, 81 (12),* pp. 2361-2370.

Sridevi, S., Rajaram, S., Parthiban, C., SibiArasan, S., & Swadhikar, C. (2011). Imputation for the analysis of missing values and prediction of time series data. *Proceedings of the Recent Trends in Information Technology (ICRTIT), 2011 International Conference on*. pp. 1158-1163.

Srinivas, K., Rao, G. R., & Govardhan, A. (2010). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. *Proceedings of the Computer Science and Education (ICCSE), 2010 5th International Conference on*. pp. 1344-1349.

Sutton-Charani, N., Destercke, S., & Denoeux, T. (2013). Learning Decision Trees from Uncertain Data with an Evidential EM Approach. *Proceedings of the Machine Learning and Applications (ICMLA), 2013 12th International Conference on*. pp. 111-116.

Tanasa, D., & Trousse, B. (2004). Advanced data preprocessing for intersites web usage mining. *Intelligent Systems, IEEE, 19 (2),* pp. 59-65.

Tang, J., Zhang, G., Wang, Y., Wang, H., & Liu, F. (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transportation Research Part C: Emerging Technologies, 51 (0),* pp. 29-40.

Tao, H., Zain, J. M., Ahmed, M. M., Abdalla, A. N., & Jing, W. (2012). A wavelet-based particle swarm optimization algorithm for digital image watermarking. *Integrated Computer-Aided Engineering, 19 (1),* pp. 81-91.

Thirukumaran, S., & Sumathi, A. (2012). Missing value imputation techniques depth survey and an imputation Algorithm to improve the efficiency of imputation. *Proceedings of the 2012 Fourth International Conference on Advanced Computing (ICoAC)*. pp. 1-5.

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology, 5 (5),* pp. 241-266.

Tsang, S., Ben, K., Yip, K. Y., Wai-Shing, H., & Sau Dan, L. (2011). Decision Trees for Uncertain Data. *Knowledge and Data Engineering, IEEE Transactions on, 23 (1),* pp. 64-78.

Twala, B., Cartwright, M., & Shepperd, M. (2005). Comparison of various methods for handling incomplete data in software engineering databases. *Proceedings* *of the 2005 International Symposium on Empirical Software Engineering, 2005*. pp. 10 pp.

Wang, D., Zhang, D., & Lu, G. (2016). A robust signal preprocessing framework for wrist pulse analysis. *Biomedical Signal Processing and Control, 23* pp. 62-75.

Wang, W., Gao, W., Wang, C., & Li, J. (2013). An Improved Algorithm for CART Based on the Rough Set Theory. *Proceedings of the Intelligent Systems (GCIS), 2013 Fourth Global Congress on*. pp. 11-15.

Yang, X.-S., & Deb, S. (2009). Cuckoo search via Lévy flights. *Proceedings of the Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*. IEEE. pp. 210-214.

Yeh, W.-C., Chang, W.-W., & Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications, 36 (4),* pp. 8204-8211.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F. *et al.* (2012). Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems, 36 (4),* pp. 2431-2448.

Yotsawat, W., & Srivihok, A. (2015). Rules Mining Based on Clustering of Inbound Tourists in Thailand. in (Eds.). *Advanced Computer and Communication Engineering Technology*. Springer. pp. 693-705.

Yusoff, M. N., & Jantan, A. (2011). Optimizing decision tree in malware classification system by using genetic algorithm. *International Journal of New Computer Architectures and their Applications (IJNCAA), 1 (3),* pp. 694-713.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8 (3),* pp. 338-353.

Zhang, C., Qin, Y., Zhu, X., Zhang, J., & Zhang, S. (2006). Clustering-based Missing Value Imputation for Data Preprocessing. *Proceedings of the Industrial Informatics, 2006 IEEE International Conference on*. pp. 1081-1086.

Zhang, S., Xindong, W., & Manlong, Z. (2010). Efficient missing data imputation for supervised learning. *Proceedings of the Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*. pp. 672-679.

Zhang, Y., Kambhampati, C., Davis, D. N., Goode, K., & Cleland, J. G. (2012). A comparative study of missing value imputation with multiclass classification for clinical heart failure data. *Proceedings of the Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*. IEEE. pp. 2840-2844.

Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G., & Lu, J. (2013). A hybrid fuzzy-based personalized recommender system for telecom products/services. *Information Sciences, 235* pp. 117-129.

Zhao, H., Guo, S., Chen, J., Shi, Q., Wang, J., Zheng, C. *et al.* (2010). Characteristic Pattern Study of Coronary Heart Disease with Blood Stasis Syndrome Based on Decision Tree. *Proceedings of the Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*. pp. 1-3.

Zhao, M., & Li, X. (2011). An application of spatial decision tree for classification of air pollution index. *Proceedings of the Geoinformatics, 2011 19th International Conference on*. pp. 1-6.