

**DATA CLUSTERING USING MIN-MIN ROUGHNESS AND ITS
APPLICATION TO CLUSTER PATIENTS SUSPECTED DIABETICS**

MOHD RIDZUAN BIN BAHARIN

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

BORANG PENGESAHAN STATUS TESIS

JUDUL **DATA CLUSTERING USING MIN-MIN ROUGHNESS AND ITS APPLICATION TO CLUSTER PATIENTS SUSPECTED DIABETICS**

SESI PENGAJIAN: **2011/2012**

Saya: **MOHD RIDZUAN BIN BAHARIN**

mengaku membenarkan tesis (Projek Sarjana Muda/Sarjana/Doktor Falsafah)* ini disimpan di Perpustakaan Universiti Malaysia Pahang dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Malaysia Pahang
2. Perpustakaan Universiti Malaysia Pahang dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila tandakan (4)

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

Disahkan oleh

(TANDATANGAN PENULIS)

(TANDATANGAN PENYELIA)

Alamat Tetap:
**63, Jln Cenderawasih,
Kpg Semerah Padi Baru.
93050 Kuching
Sarawak.**

Nama penyelia:
DR. TUTUT HERAWAN

Tarikh: _____

Tarikh: _____

- CATATAN: * Potong yang tidak berkenaan.
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT atau TERHAD.
*** Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda (PSM).

DATA CLUSTERING USING MIN-MIN ROUGHNESS AND ITS APPLICATION
TO CLUSTER PATIENTS SUSPECTED DIABETICS

MOHD RIDZUAN BIN BAHARIN

A report submitted in partial fulfillment
of the requirements for the award
of the degree of
Bachelor of Computer Science (Software Engineering)

Faculty of Computer System & Software Engineering
Universiti Malaysia Pahang

JUNE, 2012

SUPERVISOR'S DECLARATION

“I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the Degree in Computer Science (Software Engineering)”

Signature :

Supervisor : Dr. Tutut Herawan

Date :

DECLARATION

I declare that this thesis entitled “Data Clustering Using Min-Min Roughness and Its Application To Cluster Patients Suspected Diabetics” is the result of my own research except as cited in the references. The report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :

Name : Mohd Ridzuan bin Baharin

Date :

ACKNOWLEDGEMENT

Praise to Allah S.W.T, without His blessing I will not be able to complete my research. Thanks for all the strength He gave me.

A very big appreciation to my PSM supervisor, Dr. Tutut Herawan for his advices, supports and all the things that he teaches that enables me to complete this research. Even the idea of this research is his, which turn out to be a success.

Also, a very special appreciation to my fellow friends that always be along my side whenever I needed them. Muhd Noor Izzwan, Mohd Muhaymin Ali, Hafiz Kamal, Mohd Afendi Zakaria Moh Amirol Redzuan Mat Rofi, Shukri Zahari, and my other classmates. Without them I will not be able to finish this research.

Lastly, to my family that always give me their support, even when we are apart, blessing from both of my parent are the thing that keeps me going and continue on my research. Special thanks to my family, especially my parents Baharin bin Ahmad and Mistiah binti Sahmat.

ABSTRACT

In the context of information technology nowadays, there are many data exists. All of this data are scrambled over inside the computer and with the presence of internet, even more data exist. The problem with this is, when we want the needed data only, there are too many to look for and they are all scrambled over the internet databases. Therefore, there are techniques that are proposed that will provide a way to automatically mine the data and obtain only meaningful data from the huge data over the internet. The area discussed in this research is Knowledge Discovery in Databases (KDD) and the technique used is Minimum-Minimum Roughness (MMR). The dataset used will be the dataset of diabetic patients. By using this MMR technique, I intended to cluster the diabetic dataset n which each cluster will contain the data most related to each other.

ABSTRAK

Dengan adanya teknologi informasi sekarang ini, jumlah data-data semakin banyak. Data-data ini tersimpan di dalam computer-komputer dan dengan kehadiran internet, lebih banyak lagi data yang ada. Perkara ini menimbulkan masalah apabila kita inginkan data-data yang kita mahukan, tetapi data-data yg ada adalah terlalu banyak dan berselerak di internet. Oleh sebab itu, terdapat teknik-teknik yang diperkenalkan untuk mengatasi masalah ini. Bidang yang dibincangkan ialah bidang Knowledge Discovery in Databases dan teknik yang digunakan ialah teknik Minimum-Minimum Roughness. Set data yang digunakan ialah set data pesakit-pesakit diabetis. Dengan menggunakan teknik MMR ini, sy bertujuan untuk mengklusterkan data tersebut dimana setiap kluster mengandungi data-data yang berkaitan dengan satu sama lain.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
1	INTRODUCTION	1
1.1	Background	1
1.2	Problem Statement	3
1.3	Scopes	4
1.4	Objectives	4
1.5	Thesis Organization	4
2	LITERATURE REVIEW	5
2.1	Diabetes	5
2.1.1	Diabetes Description	5
2.1.2	Diabetes Symptoms	6
2.1.3	Diabetes in the World	7
2.1.4	Diabetes in Asia	8
2.1.5	Diabetes in Malaysia	9
2.1.6	Patients Suspected Diabetes	11
2.2	Knowledge Discovery in Databases	12
2.2.1	Definition of KDD	12
2.2.2	KDD Process	13
2.2.3	Examples of KDD Processes	15
2.2.4	Application of KDD in computer science	15

	fields	
2.3.	Data Mining	16
	2.3.1. Definition of DM	16
	2.3.2. Examples of DM	18
	2.3.3. Applications of DM in computer science fields	19
2.4.	Data Clustering	20
	2.4.1. Definition	20
	2.4.2. Classification vs Clustering	21
	2.4.3. Clustering Techniques	23
	2.4.4. Clustering on Numerical Dataset	24
	2.4.5. Clustering on Categorical Dataset	25
	2.4.6. Applications of Clustering Techniques	26
2.5.	Rough Set Theory	27
	2.5.1. Rough set	28
	2.5.2. Fuzzy set	29
	2.5.3. Relation between fuzzy and rough set theories	29
	2.5.4. Applications of rough set	30
2.6.	Rough Clustering	31
	2.6.1. Application of rough set in data clustering	32
	2.6.2. Rough set theory in categorical data clustering	32
3	METHODOLOGY	34
3.1.	Rough Set Theory	34
	3.1.1. Information System	35
	3.1.2. Indiscernibility Relation	39

3.1.3.	Approximation Space	40
3.1.4.	Set Approximations	40
3.2.	Min-Min Roughness	45
3.2.1.	Selecting a clustering attribute	45
3.2.2.	Model for selecting a clustering attribute	46
3.2.3.	Min-Min Roughness Technique	46
3.2.4.	Algorithm	47
3.2.5.	Example	49
3.3.	Object Splitting model	91
3.3.2.	The splitting point attributes a_4 is determined	92
3.3.3.	Cluster Purity	92
4	EXPECTED RESULTS AND DISCUSSION	93
4.1.	Datasets	93
4.1.1.	Small datasets	93
4.1.2.	Large dataset (real world dataset)	94
4.1.3.	Benchmark dataset	94
4.2.	MMR Software Development	94
4.2.1.	Interface	95
5	CONCLUSION	97
	REFERENCE	98

LIST OF TABLES

TABLE NO.	TITLE	PAGE
3.1	An information system	36
3.2	A diabetic decision system	37
3.3	Step-by-step Max-Max Roughness	46
3.4	An information system in MMR	49
3.5	Mean roughness a_1	85
3.6	Mean roughness a_2	86
3.7	Mean roughness a_3	86
3.8	Mean roughness a_4	87
3.9	Mean roughness a_5	87
3.10	Mean roughness a_6	88
3.11	Mean roughness a_7	88
3.12	Mean roughness a_8	88
3.13	Mean roughness a_9	89
3.14	Mean roughness a_{10}	90
3.15	Minimum roughness	91
3.16	MMR value	91

LIST OF FIGURES

TABLE NO.	TITLE	PAGE
1.1	KDD Process	2
2.1	Prevalence of diabetes by age group in Malaysia	10
2.2	Prevalence of diabetes in Malaysia by states	10
2.3	Overview of the steps that compose the KDD process	13
2.4	Classification	21
2.5	Clustering	22
2.6	Clustering process	23
3.1	Set approximations	42
3.2	Model for selecting a clustering attribute	46
3.3	Result of clustering	92
4.1	Start Interface	95
4.2	Calculation Interface	95

CHAPTER I

INTRODUCTION

This chapter briefly discuss on the overview of this research. It contains six parts. The first part is introduction; follow by the problem statement. Next is the motivation, followed by the scopes of the research. After that are the objectives where the research's goal is determined and lastly is the thesis organization which briefly describes the structure of this thesis.

1.1 Background

Knowledge Discovery from Databases (KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. As a branch of machine learning, KDD encompasses a number of automated methods whereby useful information is mined from data stored in databases.

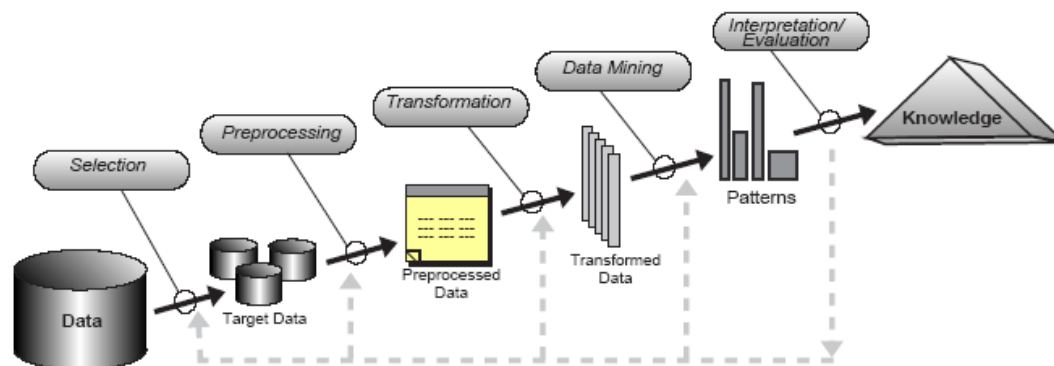


Figure 1.1: KDD Process

Data Mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. Classification and Clustering are two of the Data Mining methods. Classification involves learning a function that maps (or classifies) a data item into one of several predefined classes, while clustering involve identifying a finite set of categories or clusters to describe the data.

Diabetes is a disease in which a person has high levels of sugar in the blood. It arises when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin produced. Insulin is needed to control blood sugar. Diabetes is a chronic, potentially debilitating and often fatal disease. It is one of the oldest known diseases; it is mentioned in Egyptian manuscript from around 1550 BCE, also indentified by an Indian physician Sushruta in 6th century BCE. Diabetes appears to have been a death sentence in the ancient era and even until now, there is no cure for diabetes. Globally, diabetes ranked as the fourth leading cause of death, in terms of disease. An estimated 3.8 million people die from diabetes-related causes each year. Such causes are: heart disease, stroke, kidney disease, nerve disease, diabetic eye disease others.

The United Nations estimates the number of people globally affected by diabetes to be 246 million and approximately half of those are in India, China, Nepal and other Asian countries.

In Malaysia, diabetes is a growing concern. In 1986, a survey, namely National Health and Morbidity Survey or NHMS for short, included diabetes as a major component in

the survey. Then in the second survey, the prevalence of diabetes in Malaysia was found to be 8.2%. There was an increase in prevalence as compared to the NHMS in 1986, which only reported 6.3% in Peninsular Malaysia. Universiti Kebangsaan Malaysia's Emeritus Professor Datuk Dr Khalid Abdul Kadir said there was a "diabetic explosion" in Malaysia and wondered whether enough was being done to stop it. He said an example could be seen among the Malays in Tanjung Karang, Selangor. The prevalence was four per cent in 1984 and 6.5 per cent in 1990. Two years ago, it shot up to 20 per cent. Dr Khalid said one in seven adults in Malaysia was a diabetic. Dr Khalid attributed the growing number of diabetic cases to the lack of physical activity and excess calories accumulation as one ages. "As the population ages, we are going to see more people with diabetes," he said, adding that diabetes, hypertension and obesity seldom killed a person but they contributed to heart diseases.

Despite of the fact that diabetes is caused by both lifestyle and genetic, the lifestyle factor, or in other word; diet, contribute much to diabetes. Based on a survey made in 1996, 16.6% of adult Malaysians are facing overweight problems while 4.4% of adult Malaysians are obese. Then another survey made in 2006, shows that 29.1 Malaysians are facing overweight problem while 14% Malaysian are obese. This increase in the number of Malaysians facing overweight problems and obesity will increase the number of chronic patients as 90% of the overweight and obese are facing diabetes.

1.2 Problem Statement

There are many data clustering methods that exist, however, most of them only dealt with only numerical data type. The problem is nowadays, in real life we are dealing with categorical data type which is multi-valued data. Also, there are uncertainties in these data that need to be handled.

The clustering of this diabetics' data involves multi-valued data. Therefore, rough set technique is used because it can handle the uncertainty and also deal with the multi-valued data of the diabetics.

Having the 8.2% prevalence of diabetes found in Malaysia, the data of those diabetics should be grouped into a balanced and meaningful cluster, and in this research, based on the symptoms of the diabetics. This grouping can help in classification of the diabetics and further investigation on that disease in Malaysia.

1.3 Scopes

The scopes of this research are:

- i. The data used are from the diabetics of Hospital Ampuan Afzan.
- ii. The clustering uses min-min roughness technique.
- iii. Apply to diabetics dataset.

1.4 Objectives

There are a few objectives of this research:

- i. To partition the patients in a meaningful way based on the symptoms' closeness.
- ii. To apply the rough set clustering technique into a real life case.

1.5 Thesis Organization

The rest of this paper is organized as follows. Chapter II describes the notion of rough set. Chapter III describes the theory of rough set. Chapter IV describes the dataset, modeling process and min-min roughness data clustering. Chapter V describes the results from an application of rough set theory for clustering data and grouping diabetes patients following by discussion. Finally, the conclusion of this work is described in section 6.

CHAPTER II

LITERATURE REVIEW

This chapter briefly discusses on existing literature related with the proposed project. There are four main sections in this chapter. The first section introduces on the topic of diabetics. The second section describes some brief information on Knowledge Discovery in Databases (KDD). The third section describes Data Mining (DM) concept. The fourth section describes Data Clustering and finally a brief review of Rough Set Theory (RST) is described in the last section.

2.1. Diabetes

This section firstly presents a description and symptoms of diabetics. Further, information of diabetics in the world, Asia and Malaysia also presented. Finally, the last sub-section presents information of patient having pre-diabetes.

2.1.1. Diabetes Description

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetic,

or a person with diabetes has a condition in which the quantity of glucose in the blood is too elevated (hyperglycemia), resulting in too much glucose building up in the blood which will eventually pass out of the body in urine. Glucose in the blood gives the energy to perform daily activities while insulin allows it the glucose to move from the blood into the liver, muscle, and fat cells, providing the essential energy and growth requirements. In diabetes, glucose in the blood cannot move into cells, so it stays in the blood, which harms the cells that need the glucose for energy and also harms certain organs and tissues exposed to the high level glucose. There are three types of diabetes; type 1 diabetes or insulin-dependent diabetes, is an auto-immune disease where the body's immune system destroys the insulin-producing beta cells in the pancreas; type 2 diabetes or non-insulin dependent diabetes, the most common form of diabetes, is characterized by insulin resistance and relative insulin deficiency, strongly genetic in origin but lifestyle factors such as excess weight, inactivity, high blood pressure and poor diet are major risk factors for its development; and gestational diabetes mellitus (GDM) or carbohydrate intolerance, usually developed during pregnancy and usually the carbohydrate intolerance returns to normal after the birth but the mother has a significant risk of developing permanent diabetes while the baby is more likely to develop obesity and impaired glucose tolerance and/or diabetes later in life.

2.1.2. Diabetes Symptoms

These are some of the common early warning signs of diabetes: excessive thirst that is unrelated to exercise, hot weather or short-term illness; excessive hunger, even after eaten; frequent urination; fatigue or tiredness possibly severe enough to make you fall asleep unexpectedly after meals; and sudden weight loss or dramatic change in weight. While many of the signs and symptoms of diabetes can also be related to other causes, testing for diabetes is very easy, and the constant/regular presence of these symptoms over an extended period of time should be cause to visit a doctor. Also, there are minor and less recognizable symptoms of diabetes, which are; blurry vision, occur because diabetes can lead to macular degeneration and eventual blindness; numbness, or tingling in the hands and feet may occur due to peripheral neuropathy, a symptom of diabetes,

causes nerve damage in the extremities; slow healing wounds, result of diabetes-related impaired immune system function; recurrent or hard-to-treat yeast infections in woman, another sign of impaired immune function; and dry or itchy skin, may result from peripheral neuropathy which affects circulation and proper sweat gland function.

2.1.3. Diabetes in the world

According to the report of diabetes published by WHO (World Health Organization) in 2009, there were at least 220 million diabetics, and the WHO also estimated that there will be at least 336 million diabetics in the world in 2030. There are 5% of global deaths which are caused from diabetes complications. The WHO had stated a few facts about diabetes:

- i. There are more than 346 million people worldwide have diabetes. There is an emerging global epidemic of diabetes that can be traced back to rapid increases in overweight, obesity and physical inactivity.
- ii. Diabetes is predicted to become the seventh leading cause of death in the world by the year 2030. Total deaths from diabetes are projected to rise by more than 50% in the next 10 years.
- iii. Type 2 diabetes is much more common than type 1 diabetes. Type 2 accounts for around 90% of all diabetes worldwide. Reports of type 2 diabetes in children, which previously rare, have increased worldwide. In some countries, it accounts for almost half of newly diagnosed cases in children and adolescents.
- iv. Cardiovascular disease is responsible for between 50% and 80% of deaths in people with diabetes. Diabetes has become one of the major causes of premature illness and death in most countries, mainly through the increased risk of cardiovascular disease (CVD).
- v. In 2004, an estimated 3,4 million people died from consequences of high blood sugar.

- vi. 80% of diabetes deaths occur in low and middle income countries. In developed countries most people with diabetes are above the age of retirement, whereas in developing countries the most frequently affected are aged between 35 and 64.
- vii. Diabetes is a leading cause of blindness, amputation and kidney failure due to lack of awareness about diabetes, combined with insufficient access to health services and essential medicines.

2.1.4. Diabetes in Asia

Research published in the medical journal *Lancet* reveals that life-threatening diabetes is becoming an epidemic not only in North America, but in Asia as well and it appears to be only getting worse. According to doctors at the Catholic University of Korea in Seoul, 194 million Asians were diabetic in 2003, a statistic that could soar to 330 million by the year 2025. The *Lancet* research suggests Asians are developing diabetes at younger age and at lower weight, they suffer longer complications and they also die earlier than people in developed countries. This onset of adult diabetes in increasingly younger populations will negatively affect Asian countries economically, as a result of higher health costs and mortality rates. According to the International Diabetes Federation (IDF)'s 2003 statistics, the top 5 countries with the largest number of diabetics were: India with 35.5 million diabetics; China with 23.8 million diabetics; USA with 16.0 million diabetics; Russia with 9.7 million diabetics; and Japan with 6.7 million diabetics. Also, there are four more countries in Asia among the top ten countries with highest number of diabetics; Indonesia, Pakistan, Bangladesh, and Philippines [9]. The WHO and the IDF predict that the number of diabetics on Asia could increase to 160 million by year 2025. Conservative estimates based on population growth and ageing and rate of urbanization in Asia show that India and China will remain the two countries with the highest numbers of people with diabetes by 2030. The current estimated age-adjusted prevalence of diabetes in China (4.2%) is expected to increase by 1.9% to 5.0% in 2030. However, results from a survey conducted in 2007-08 in China reported a higher prevalence (9.7%, including previously undiagnosed

diabetes). A similar increase is expected in India (from 7.8% in 2010 to 9.3 in 2030). Southern and Eastern Asian countries account for half the deaths attributable to diabetes worldwide and the consequences of a rising incidence and prevalence of diabetes and other chronic diseases will be particularly important, both locally and globally. A concurrent increase in the prevalence of overweight and obesity over the same time period has been observed in many Asian countries. In rural China, the prevalence of overweight has increased from 5.3% in men and 9.8% in women in 1992 to 13.6% in men and 14.4% in women in 2002. Corresponding figures for obesity were 0.5% in men and 0.7% in women in 1992 and 1.8% in men and 3.0% in women in 2003. This trend has profound implications for the expected number of diabetes patients who will be diagnosed in this region for future decades. Type 2 is due primarily to lifestyle factors and genetics and most of Asian patients have a first-degree relative with diabetes. Also, most of the loci originally associated with diabetes in European populations have been replicated in Asian population. The urbanization and migration of Asians, which is expected to increase, would be the cause of the rise in the global prevalence of diabetes.

2.1.5. Diabetes in Malaysia

Studies show that the prevalence of type 2 diabetes is escalating at a phenomenal scale and very likely we are heading towards epidemic proportions. Malaysia is too to be affected by this as there are major shifts in the lifestyles and longevity of the population. In other words, Malaysia has the right ingredients to set the scene for the explosion of diabetes [m]. Professor Datuk Dr Khalid Abdul Kadir, University Kebangsaan Malaysia Emeritus and also a professor of medicine at Monash University, said there was a “diabetic explosion” in Malaysia and wondered whether enough was being done to stop it. He said one in seven adults in Malaysia was a diabetic and more people below the age of 45 are getting diabetes. He also said that with modernization and economic progress, there would be an explosion of “metabolic catastrophe” in Asia, including Malaysia, due to obesity, hypertension and diabetes. According to him, in 1990, the prevalence of obesity and diabetes among the Orang Asli, the hunter-gatherers in the jungle fringes of Pahang or in settlements at Carey Islands and Ulu Langat outside Kuala Lumpur was

zero. But over five years, the Institute for Medical Research found 5% of the resettled Orang Asli had diabetes. He attributes the growing number of diabetic cases to the lack of physical activity and excess calories accumulation as one ages. Statistics pointed that Malaysia had the fourth highest number of diabetes cases in Asia, with 800,000 in 2007. In the Malaysian Burden of Disease and Injury Study, it was estimated that for year 2000, there were 2,261 deaths attributed to diabetes in which 847 of them are men and 1404 are women. The earliest diabetes studies carried out in Malaysia were in 1960 and in 1966. The first National Health and Morbidity Survey (NHMS) in Malaysia was carried out in 1986 where prevalence of diabetes among adults of age 35 years old and above was found to 6.3%. Then 10 years later, in NHMS II, the figure had increased by one third to 8.3% among adults of age 30 years and above. This shocking rise spurred the initiation of numerous national healthy lifestyle campaigns by the Ministry of Health Malaysia. A national steering committee was set up to improve the screening and management of diabetes in primary and secondary care clinics. In 2006, the third NHMS is conducted. The result shows that diabetes is also detected in the younger age group, between the ages of 18 to 30 years old. Also, there was a general increasing trend in diabetes prevalence with age; from 2.0% in the 18-19 years old age group to a prevalence ranging between 20.8 to 26.2% among the 60-64 years old shown in Figure 1.

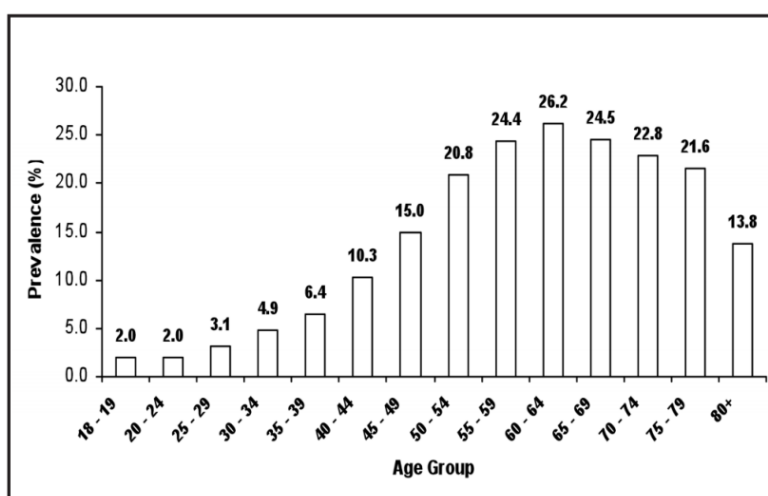


Figure 2.1: Prevalence of diabetes by age group in Malaysia

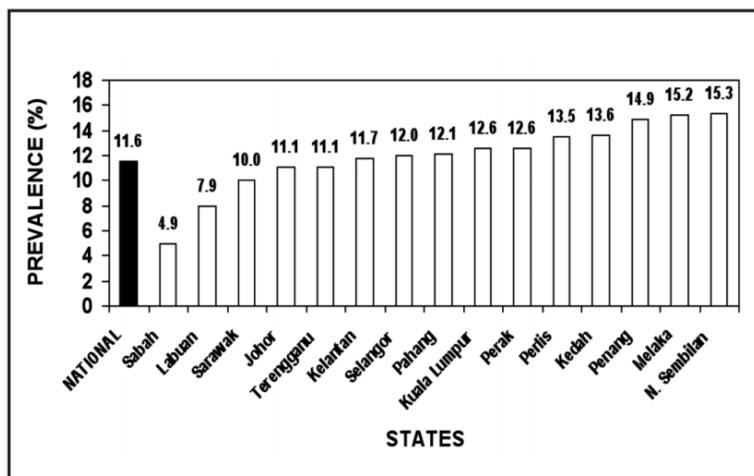


Figure 2.2: Prevalence of diabetes in Malaysia by states

Among the states, Negeri Sembilan, Malacca and Penang had the highest prevalence of diabetes at 15.3%, 15.2% and 14.9% respectively as shown in Figure 2. The prevalence was higher in the urban at 12.2% compared to the rural areas at 10.6%. No significant gender difference was observed; 11.9% in the males while 11.3% in the females.

2.1.6. Patient suspected Diabetes

Patient suspected diabetes can be also said as people having pre-diabetes or people with the risk of diabetes. Pre-diabetes is the state that occurs when a person's blood glucose levels are higher than normal but not high enough for a diagnosis of diabetes. This means that this people have the risk of getting diabetes. According to the American diabetes association, in the Diabetes Prevention Program, 11% of people with pre-diabetes developed type 2 diabetes each year during the average 3 years of follow up. Other studies show that many people with pre-diabetes develop type 2 diabetes in 10 years. People suspected diabetes might not know that they are having diabetes risk. In fact, many people that have diabetes do not realize it because of the symptoms develop so gradually, people often do not recognize them. Some people have no symptom at all.

There are many risk factors for type 2 diabetes, as shown below:

- i. Obesity; the National Center for Health Statistics states that 30% of adults are obese. That is 60 million people. Greater weight means a higher risk of insulin resistance, because fat interferes with the body's ability to use insulin.
- ii. Sedentary lifestyle; muscle cells have more insulin receptor than fat cells, so a person can decrease insulin resistance by exercising. Being more active also lowers blood sugar levels by helping insulin to be more effective.
- iii. Unhealthy eating habits; 90% of people who have been diagnosed with type 2 diabetes are overweight. Unhealthy eating contributes largely to obesity. Too much fat, not enough fiber and too many simple carbohydrates all contribute to a diagnosis of diabetes.
- iv. Family history and Genetics; it appears that people who have family members who have been diagnosed with type 2 diabetes are at a greater risk for developing it themselves. Lifestyle plays an important part in determining who gets diabetes.
- v. Increased age; the older we get, the higher our risk of type 2 diabetes.
- vi. High blood pressure and high cholesterol; having metabolic syndrome increases the risk of heart disease, stroke and diabetes.
- vii. History of gestational diabetes; gestational diabetes affects 4% of all pregnant women. Many women who have gestational diabetes develop type 2 diabetes years later. Their babies are also at some risk for developing diabetes in later in life.

2.2. Knowledge Discovery in Databases

This section presents a definition, processes, examples and applications of Knowledge Discovery in Databases (KDD).

2.2.1. Definition of KDD

The rapidly growing volume and complexity of modern databases make the need for technology to describe and summarize the information they contain increasingly important. Knowledge Discovery in Databases (KDD) and data mining are new research areas that try to deal with this problem. KDD is defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. There are also many other terms, appearing in some articles and documents, carrying a similar or slightly different meaning, such as knowledge mining from database, knowledge extraction, data archeology, data grudging, data analysis and so on. The goal of KDD is to make the patterns of data understandable to humans. The data are a set of facts and pattern is an expression in some language describing a subset of the data or a model applicable to the subset. The discovered patterns should be valid on new data with some degree of certainty. The patterns need to be novel and potentially useful, that is lead to some benefit to the user or task. Having KDD means that the interesting knowledge, regularities or high level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large database thereby serves as rich and reliable sources for knowledge generation and verification.

2.2.2. KDD Processes

The KDD process involves using the database along with any required selection, preprocessing, subsampling and transformations of it; applying data mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The overall KDD process includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge.

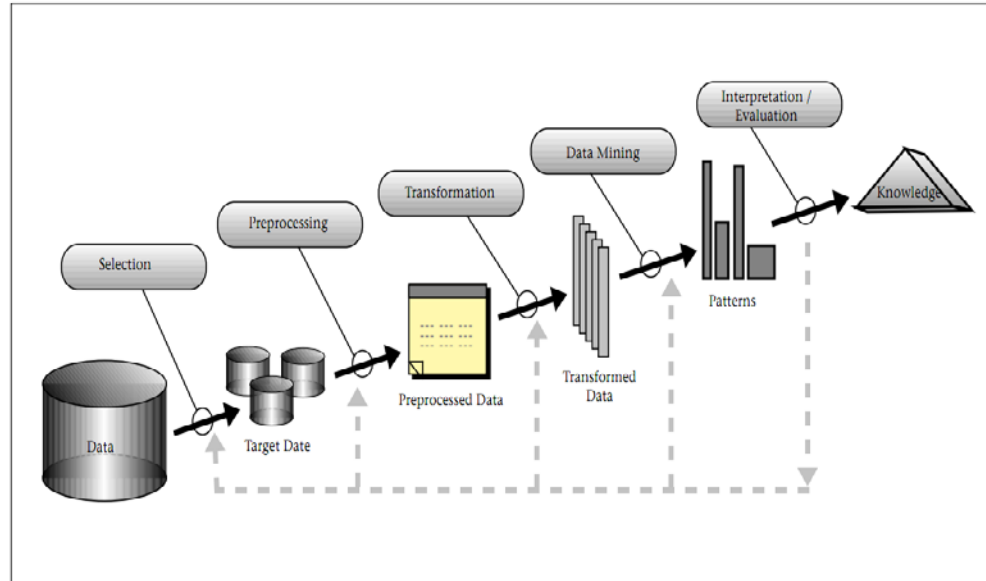


Figure 2.3: Overview of the steps that compose the KDD process

The process of KDD consists of the following steps:

- i. Developing an understanding of the application domain, the relevant prior knowledge, and the goal(s) of the end user.
- ii. Creating or selecting a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- iii. Data cleaning and preprocessing: this step includes, removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.
- iv. Data reduction and projection: finding useful features to represent the data depending in the goal of the task. This may include dimensionality reduction or transformation to reduce the effective number of variables under consideration or to find the invariant representations of the data.
- v. Matching the goals to a particular data mining method such as summarization, classification, regression, clustering etc. Model and

hypothesis selection, choosing the data mining algorithm(s) and methods to be used for searching for data patterns.

- vi. Exploratory analysis and model and hypothesis selection: choosing the data mining algorithms(s) and selecting method(s) to be used for searching for the data patterns. This process includes deciding which models and parameters might be appropriate and matching particular data mining method with the overall criteria of the KDD process.
- vii. Data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data mining method by correctly performing the preceding steps.
- viii. Interpreting mined patterns: possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.
- ix. Acting on discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

2.2.3. Examples of KDD Processes

One example of a system that applies KDD in astronomy area is SKICAT, a system used by astronomers to perform image analysis, classification and cataloging of sky objects from sky-survey images. In its first application, the system was used to process 3 terabytes (10^{12} bytes) of image data resulting from the Second Palomar Observatory Sky Survey, where it is estimated that on the order of 10^9 sky objects are detectable.

SKICAT can outperform humans and traditional computational techniques in classifying faint sky objects.

While in marketing, the primary application is database marketing systems, which analyze customer databases to identify different customer groups and forecast their behavior. Another notable marketing application is market-basket analysis system, which find patterns such as, “If customer bought X, he/she is also likely to buy Y and Z”. Such patterns are valuable to retailers.

KDD is also applied in fraud detection system such as HNC Falcon and Nestor PRISM systems which are used for monitoring credit card fraud, watching over millions of accounts. Also, the FAIS system from the U.S. Treasury Financial Crimes Enforcement Network, is used to identify financial transactions that might indicate money laundering activity.

Lastly, a novel and increasingly important type of discovery is one based on the use of intelligent agents to navigate through an information-rich environment. These systems ask the user to specify a profile of interest and search for related information among a wide variety of public domain and proprietary sources.

2.2.4. Application of KDD in computer science fields

Computer science is the study of the theoretical foundations of information and computation and of practical techniques for their implementation and application in computer system. In today’s society, the Information Technology (IT) is an increasingly part of all economic, technological, educational and even cultural sectors. Through applications such as e-commerce, networking, and digital administration, the IT evolution has become one of the most important factors in shaping the future of our social system. Currently, many kinds of data are being generated and stored about all kind of human endeavors for example like the widespread use of the bar codes for most commercial products, the computerization of many business and government transaction, and the advances in data collection tools have provided us with huge amounts of data. These data are stored or recorded in the form of computer databases, where the computer technology can easily access it. Millions of databases have been

used in business management, government administration, scientific and engineering data management, and many other applications. It is noted that the number of such databases keeps growing rapidly because of the availability of powerful and affordable database systems. However, this very large amount of data created problem on how to extract from them useful, task-oriented knowledge. This is where the KDD knowledge, together with computer science knowledge is applied to extract the useful data. Using the computer technology itself to access the large computer database, a program, or software that contains the needed steps and algorithms can be developed for the automation of data extraction, to transform effectively the increasing amount of data in valuable knowledge and to discover the hidden relationships within the data. This requirement has led to the development in the computer science discipline of the data mining field, which combines methods and approaches from the fields of machine learning, databases, cluster analysis, statistics and visualizations.

2.3. Data Mining

This section presents a definition, examples and applications of Data Mining (DM).

2.3.1. Definition of DM

The term data mining is always been refer to as the same as KDD, or many other terms such as knowledge extraction, information discovery, information harvesting, data archeology, data pattern processing, knowledge mining from databases, data analysis and many more that brings the meaning as to find or extract the useful patterns of data. However, KDD is the overall process of discovering the useful knowledge from data, while data mining of one of the steps in KDD. Data mining is seen as the key element in KDD. It provides a new thought for organizing and managing tremendous data. Data mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data. The data mining component of the KDD process often involves repeated iterative application of particular data-mining methods. Data mining

uses automated tools employing sophisticated algorithms to discover hidden patterns, associations, anomalies and/or structure from large amounts of data stored in data warehouses or other information repositories. Data mining tasks can be descriptive, for example like discovering new patterns describing the data, and predictive for example like predicting the behavior of the model based on available data. Data mining involves fitting models or determining patterns from observed data. The fitted models play the role of inferred knowledge. Deciding whether the model reflects useful knowledge or not, is a part of the overall KDD process for which subjective human judgment is usually required.

Typically, a data mining algorithm constitutes some combination of the following three components:

- i. The model: the function of the model (such as classification and clustering) and its representational form (such as linear discriminants, neural networks). A model contains parameters that are to be determined from the data.
- ii. The preference criterion: A basis for preference of one model or set of parameters over another, depending in the given data. The criterion is usually some form of goodness-of-fit function of the model to the data, perhaps tempered by a smoothing term to avoid over fitting, or generating a model with too many degrees of freedom to be constrained by the given data.
- iii. The search algorithm: the specification of an algorithm for finding particular models and parameters, given the data, model(s) and preference criterion.

A particular data mining algorithm is usually an instantiation of the model/preference/search components. The more common model functions in current data mining practice include the following:

- i. Classification: classifies a data item into one of several predefined categorical classes.
- ii. Regression: maps a data item to a real valued prediction variable.
- iii. Clustering: maps a data item into one of several clusters, where clusters are natural groupings of data items based in similarity metrics or probability density models.
- iv. Rule generation: extracts classification rules from the data.
- v. Discovering association rules: describes association relationship among different attributes.
- vi. Summarization: provides a compact description for a subset of data.
- vii. Dependency modeling: describes significant dependencies among variables.
- viii. Sequence analysis: models sequential patterns. Like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time.

2.3.2. Examples of DM

In the business and investment area, numerous companies use data mining for investment. Most companies do not describe their systems except for LBS Capital Management. Its system uses expert system, neural nets and genetic algorithms to manage portfolios totaling \$600 million; since its start in 1993, the system has outperformed the broad stock market.

Another example of data mining application is DBMiner, a system for data mining in relational databases and data warehouses. DBMiner has been developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses. It implements a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction and clustering. DBMiner incorporates several interesting data mining techniques including OLAP (online analytical processing) and attribute-oriented induction, statistical analysis, progressive deepening for mining multiple-level knowledge, and meta-rule guided mining, which provides a

user-friendly, interactive data mining environment with good performance. This system has been tested on several medium to large databases, including NSERC (Natural Science and Engineering Research Council of Canada) research grant information system, and U.S. City-County Data Book, with satisfactory performance.

2.3.3. Applications of DM in computer science fields

Data mining or Knowledge Discovery in databases involves extracting useful information from large databases. This is due to the increasing information of the things in the world that need to be recorded day by day. Most of the times, with the advancement of technologies that we have achieved until now, the data are being stored in the form of computer databases. Due to this existence of vast amount of data, there is a need for the conversion of it into useful information. There are many different methods have been developed to address this need and such methods have come to be known as data mining. These methods enables the creation of algorithms to be applied into computer science knowledge to form application that can perform data mining process, to help in obtaining the useful information from the large databases.

Data mining has been applied to a variety of business problems, such as Supply Chain Management (SCM), Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM). Considerable amount of work has been done in CRM to apply data mining in cooperation with Artificial Intelligence (AI) to market based analysis and customer segmentation. Symeonidis et al. (2003) integrate data mining techniques such as clustering with agent technology to identify SCM and CRM patterns, thus facilitate ERP implementation. Chen et al. (2005) apply association rule mining for distribution centers to determine the order batches. Aitken et al. (2003) propose a classification system to classify products and therefore develop appropriate supply chain strategies. Srinivasan et al. (1999) use a clustering algorithm to evaluate inventory decisions. Hong et al. (2005) apply a clustering technique to group customers which can then be used for supplier selection.

One example of data mining can be seen in the paper by Show-Jane Yen and Yue-Shi Lee. Their paper is about an incremental data mining algorithm to discover web access

pattern. This involves the discovery and analysis of useful information from the web, which is also called as web mining. They stated the problem: that most of the web systems only statically provide the homepages about the page contents and information services, with no consideration of the behaviors of the users and because of this, the users spend a lot of times to search wait for the information that they needed. Therefore, data mining applied to help in finding the useful traversal paths for the users in the web system. This is to discover the paths traversed by a sufficient number of users from the web logs, which can be used for prefetching and suggestion for the web users, thus helping them to look for needed information faster.

2.4. Data Clustering

This section presents a clustering definition, its comparison with classification, clustering techniques examples, application of clustering techniques, and clustering on numerical and categorical datasets.

2.4.1. Definition

Cluster is defined as a collection of data objects which are similar to one another within the same cluster but dissimilar to the objects in the other clusters. A simple definition of data clustering is the grouping together of similar data item into clusters. A more elaborated definition of data clustering: these clusters should reflect some mechanism at work in the domain from which instances or data points are drawn, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances.

To provide more understanding of clustering, let's take an example, the search for World Wide Web pages on a given topic: "fly". This keyword will be send to the search engine and the results are returned. Results that contain the instances of the keyword "fly" will appear and for more sophisticated search engines, they may extend the search result by including the derivation of the keyword "fly" such as flying or "flies". This is because of

the idea of the word “flying” belong to the same semantic group as “fly” and also, similar to “flies”, it is a grammatical transformation of the word “fly”. Notice that “fly” can refer to the action of an aircraft flying and also the insect fly. Both are equally valid and such words are known as homographs. Fortunately, clustering algorithms, allow unlabelled text documents to be placed in each of these two sets. It attempts to identify the groupings of the results, which are similar to some respect. In the first cluster should have the results of which referred to the act of flying and the second cluster would contain the result of which referred to the insect fly. Clustering works because these search results refer to the insect are more likely to contain other words that are meaningful to a discussion about the insect, and similar for the search result that refer to the act of flying, the other words are meaningful to the discussion about the act of flying.

2.4.2. Classification vs Clustering

Classification is defined as the task to learn to assign instances to predefined classes while clustering is defined as the task to learn a classification from the data. This means that, in classification, no predefined classification is required. Both of them are the central concept of pattern recognition, and both of them are used as important knowledge discovery tools in modern machine learning process. Classification involves assigning input data into one or more pre-specified classes based on extraction of significant features or attributes and the processing or analysis of these attributes. Classification requires supervised learning, or also called as learning by example. Supervised learning is a process where the system attempts to find concept descriptions for classes that are together with pre-classified examples. It assumes that the data are labeled or defined. Figure 3 explains the classification where the data or observation (O) is mapped to the classes (C) with a predefined patterns or descriptions (P).

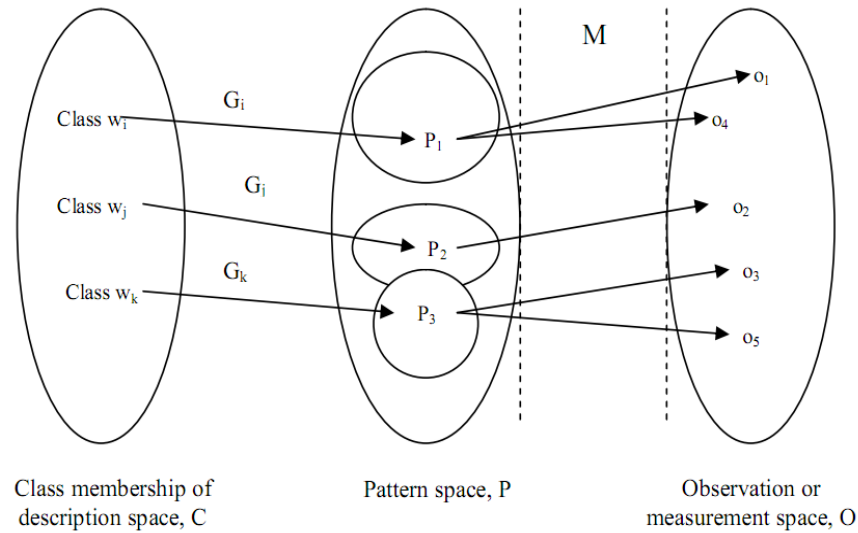


Figure 2.4: Classification

However, in many cases, there are no clear structures in the data where the classes are not defined. There are no obvious number of classes and the relationship between the data's attributes or characteristics are not clear. Besides, when dealing with everyday's data, over time, the characteristics or attributes of the class-specific pattern can change continuously. Therefore, clustering is used to solve this. Clustering involves automatic identification of groups of similar objects or patterns. This is done by maximizing the inter-group similarity and minimizing the intra-group similarity, and as a result, a number of clusters would form on the measurement or observation space. Then the data can be easily recognized and be assigned to the clusters suitable label or feature description. This is shown in the figure below.

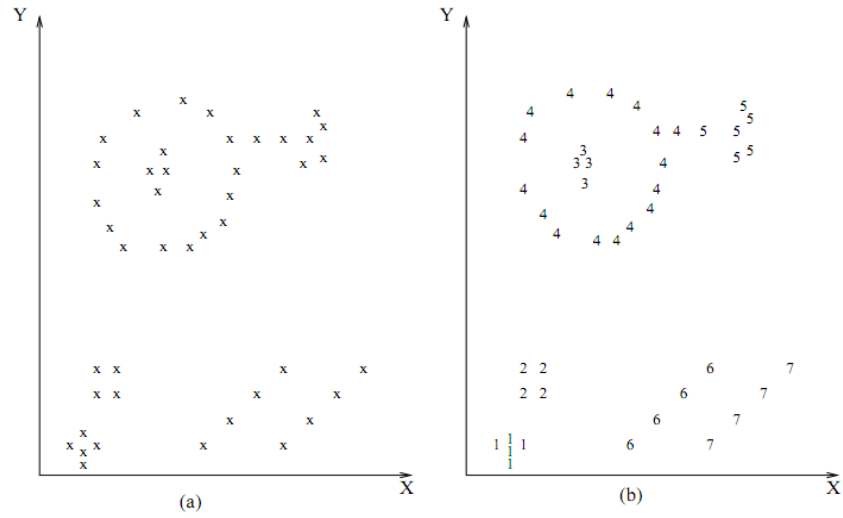


Figure 2.5: Clustering

Clustering requires unsupervised data learning where the task is only directed to search the data for interesting associations, and attempts to group elements by postulating class descriptions for sufficient many classes to cover all items in the data.

2.4.3. Clustering Techniques

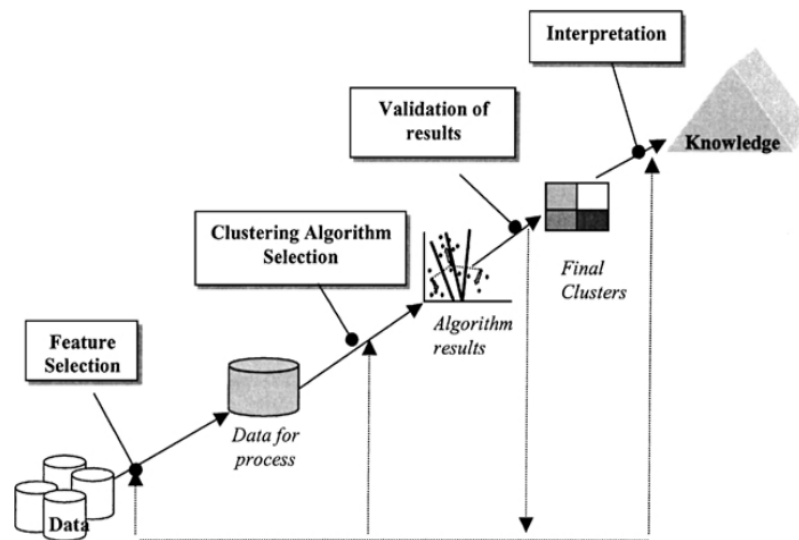


Figure 2.6: Clustering process

The clustering techniques involve the following process:

- i. Feature selection: the goal of this process is to select the proper features on which the clustering is to be performed so that as many as possible information can be obtained concerning the task of the interest. Therefore, the preprocessing of the data may be necessary prior to the utilization in clustering task.
- ii. Clustering algorithm. This step refers to the choice of an algorithm which results in the definition of a good clustering scheme for a data set. A proximity measure and a clustering criterion mainly characterize a clustering algorithm as well as its efficiency to define a clustering scheme that fits the data set.
 - o Proximity measure is a measure that quantifies how similar two data points are or in other words, how strong the relationship between the two data. In most of the cases, we have to ensure that all selected features contribute equally to the computation of the proximity measure and there are no features that dominate others.

- Clustering criterion. In this step, the clustering criterion is to be defined. It can be expressed via a cost function or some other type of rules. The type of clusters that are expected to occur in the dataset must be taken into account. Then, a suitable clustering criterion can be defined, and thus, leads to a partitioning that fits the data set well.
- iii. Validation of the result. The correctness of the clustering algorithm results is verified using appropriate criteria and techniques. Since the clustering clusters are not predefined, the final partition of the data requires some kind of evaluation in most applications.
- iv. Interpretation of the results. In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analysis in order to draw the right conclusion.

2.4.4. Clustering on Numerical Dataset

Data is divided into two types, numerical data and categorical data. Numerical data, or also called as quantitative data, is defined as data that consists of numerical measures or counts. With numerical data values, it makes sense to calculate meaningful measure of center, or to examine the spread of the data. Clustering numerical data relies on a metric that determines the distance of data pairs or the level of similarity of the data pairs. The main metrics used are Euclidean distance, the Canberra Metric, the correlation coefficient and the Mahalanobis distance. From these mentioned metrics, two different approaches can be defined: hierarchical and non- hierarchical clustering.

i. Hierarchical

The data are not partitioned into pre-defined set of clusters but are linked to the nearest group forming a single cluster containing all objects (agglomerative methods) or divided in up to n number of clusters, each containing a single object

(divisive methods). The groups are joined in a dendrogram that shows the similarity structure of the data and the number of groups generated depends on a cut parameter based on the user's analysis of the dendrogram generated.

i. Non-hierarchical.

K-means is the most common non-hierarchical algorithm. It starts with an initial partition of the cases into k number of clusters and iterates in order to find the best clusters, which is the one that minimizes the intra-clusters instances). K-means algorithm is a very fast algorithm that has the drawback on relying on the pre-definition of the number of clusters by the users. This is a problem because sometimes, the knowledge of the ideal number of clusters is not available.

The Fuzzy c-Means algorithm is a fuzzy clustering algorithm used to establish and optimal classification for the data. It is a generalization of the K-Means algorithm that makes the class membership to become a relative one, allowing an object to belong to several classes at the same time with different degrees.

2.4.5. Clustering on Categorical Dataset

Categorical data, or also known as qualitative data, consists of attributes, labels or non-numerical categories. In dealing with categorical data, each subject of the study is placed into a category, such as types or sizes. This rises a problem to cluster the data, where the domains of the individual attributes are discrete valued and not naturally ordered. The challenges to cluster the categorical type of data rise due to; the lack of natural order on the individual domains, which makes a large number of traditional similarity measures ineffective; the high dimensionality of the datasets; and many categorical datasets do not exhibit clusters over the full set of dimensions.

Previously, most of the clustering algorithms focus on numerical data. However, nowadays, the data mining community is inundated with a large collection of categorical data like those collected from banks, health sectors, web-log data, biological data, market retailing data and many others. Banking sector or health sector data are primarily mixed data containing numeric attributes like age, salary, etc. and categorical data like

sex, smoking or non-smoking, etc. Even most of the data existed in the databases nowadays are categorical, where the attribute values cannot be naturally ordered as numerical values. A few algorithms have been proposed in recent years for clustering categorical data. The K-means algorithm suggested by Mac Queen only works on numeric data. Then, the K-modes presented by Huang has expanded the previous K-means algorithm to deal with the categorical data. The fuzzy K-modes algorithm put forward by Huang and NG has improved the clustering accuracy by fuzzy processing technology. However, those algorithms have a shortcoming that is the clustering results are often dependent on the selection of the initial points. Then, with an aim at the shortcoming, Bradley and Fayyad posed the refining initial points for k-means clustering; Sun, Zhu and Chen advanced the iterative initial-points refinement algorithm for categorical data clustering; and D. W. Kim, K. H. Lee and D. Lee suggested the fuzzy clustering of categorical data using fuzzy centroids. Concept clustering algorithm is another kind of methods to deal with the clustering of categorical data. This kind of method not only can realize the clustering process, but provide the concept descriptions of clusters as well. The hierarchical clustering algorithm can deal with both numerical and categorical data. ROCK algorithm is a type of agglomerative hierarchical clustering algorithm for categorical data. Rough set theory (RST) which is suggested by Pawlak is a new mathematical tool to deal with vagueness and uncertainty, with successful applicable results obtained in many fields of information system. Recently, RST has displayed the vast applicable future in KDD field.

2.4.6. Applications of Clustering Techniques

One example of application that uses clustering technique is the CASSIOPEE troubleshooting system, which is developed as part of a joint venture between General Electric and SNECMA, was applied by three major European airlines to diagnose and predict problems for the Boeing 737. To derive families of faults, clustering methods are used. CASSIOPEE received the European first prize for innovative application. Clustering procedures have been applied to a wide variety of topic and areas. Uses of clustering techniques can be found in pattern recognition, compression, classification

and classic disciplines as psychology and business. They are useful for deriving hierarchies of plants in biology, for classifying individuals into personality types in psychology, for classifying stars or optical pictures of galaxies in astronomy and many more. This makes clustering a technique that merges and combines techniques from different disciplines such as mathematics, physics, math-programming, statistics, computer sciences, artificial intelligence and databases among others.

Below are some more of the applications and descriptions of clustering in various fields:

- i. In business, clustering may help marketers discover significant groups in their customers' database and characterize them based on the customers' patterns of purchasing.
- ii. Biology. Data clustering can be used to define taxonomies, to categorize genes with similar functionality and gain insights onto structures inherent in populations.
- iii. Spatial data analysis. Spatial data or also known as geospatial data or geographic information is the data that identifies the geographic location of features and boundaries on Earth. There are huge amounts of spatial data, obtained from satellite images, medical equipment, Geographical Information Systems (GIS), image database exploration and others. Because of too many data, it is difficult to examine it. Therefore, clustering can help to automate the process of analyzing and understanding the data by identifying and extracting interesting characteristics and patterns that may exist in the spatial database.
- iv. Web mining. The web consists of a huge collection of data. Clustering help to discover the significant groups of data which is useful n information discovery.

2.5. Rough Set Theory

This section presents the history of rough set theory, the definition of fuzzy set and its relation with rough set theory. In the last sub-section, the applications of rough set theory are presented.

2.5.1. Rough set

Rough set have been introduced as a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications. This theory is introduced by Pawlak (1981) and becomes popular among scientist around the world due to its fundamental importance in the field of artificial intelligence and cognitive sciences.

The building idea of set theory is an assumption that every set of the universe of discourse, we associate some information in the form of data and knowledge. It assumes that an object is characterized with features. Objects clustered by the same information are similar with respect to the available information about them. The similarity generated this way forms the basis for rough set theory. Objects described by identical data are indiscernible. Therefore, rough set theory is basically about interpreting, characterizing, representing and processing the indiscernibility of the data. It provides systematic method for representing and processing vague concepts caused by indiscernability in situations with incomplete information or lack of knowledge. In rough set theory, a set of all similar objects is called elementary and it makes a fundamental atom of knowledge. Any union of elementary sets is referred to as a precise set as opposed to an imprecise (rough) set. As a result, each rough set has boundary-line objects. This means that some objects cannot be classified for sure as members of the set or its complement. In other words, when the available knowledge is employed, boundary-line cases cannot be properly classified. Therefore, rough sets can be considered as uncertain. In rough set theory, a set can be defined using a pair of crisps sets, lower and upper approximations of a set. Incorporating classic axioms with rough sets theory, we have rough set based logics and new features that are very useful in intelligent decision making. Based on uncertain and inconsistent data, rough set logic allows correct reasoning and discovery of hidden associations.

The basic idea of rough sets is that some cases may be clearly labeled in a set X (positive region) while some cases may be clearly labeled as not being in set X (negative region) but limited information prevents the labeling of all possible cases clearly. The remaining cases cannot be distinguished and lie at the “boundary region”. Rough set theory calls the positive region as “lower approximation” of set X that yields no false positives. The

positive region plus the boundary region make up an “upper approximation” of set X that yields no false negatives. The lower approximation of a concept consists of all objects that definitely belong to the concept. The upper approximations of the concept consist of all objects that possibly belong to the concept.

2.5.2. Fuzzy set

Fuzzy set is a class of objects with a continuum of grades of membership. Such a set is characterized by a membership (characteristics) function which assigns to each object a grade of membership ranging between zero and one. The notions of inclusion, union, intersection, complement, relation, convexity, etc., are extended to such sets, and various properties of these notions in the context of fuzzy sets are established thus allowing partial memberships. In particular, a separation theorem for convex fuzzy sets is proved without requiring that the fuzzy sets be disjoint. In other words, fuzzy set is defined by a membership function from a universe U to the unit Interval. This introduces generalized notions of sets and members of sets, compared with classical a set which is a two-valued logic. The membership values may be interpreted in terms of truth values of certain propositions, and fuzzy set operators in terms of logic connectives in many valued logic. The theory of fuzzy sets may be viewed as the deviation of the classical set theory, in which both are having the same vocabulary, but different interpretation of the vocabulary.

2.5.3. Relation between fuzzy and rough set theories

Theories of fuzzy sets and rough sets are generalizations of classical set theory for modeling vagueness and uncertainty. It is generally accepted that they are related but distinct and complementary theories even though some authors argues that they are the same theory, only one is more general from the other. Rough set theory deals with the indiscernibility between objects. The fuzzy set theory deals with the ill-definition of the boundary class through a continuous generalization of set characteristic functions. The indiscernability of objects is not used in fuzzy set theory. A fuzzy set may be viewed as

a class with unsharp boundaries, while a rough set is a crisp set which is coarsely described.

A fuzzy set is defined as a membership function from a universe U to the unit interval. This introduces generalized notions of sets and members of sets, compared with classical sets. Meanings of classical sets and set-theoretic operators have to be modified to accommodate these notions. The set-oriented view of rough set starts from classical set algebra and associates a fuzzy set with each subset of the universe. Vagueness in concept formation and representation comes from our inability to describe a precisely defined concept in situations with incomplete information. This model captures another source of vagueness. Rough membership functions may be interpreted as special type of fuzzy membership functions, which can be interpreted in terms of probabilities defined simply by cardinalities of sets. In general, one may use a probability function on U to define rough membership functions. Given these ideas of fuzzy sets and rough sets, both of the theory can be viewed as the deviation of the classical set theory.

Two views of rough set theory provide distinct generalizations of classical set theory, namely; deviation and extension. By using an equivalence relation on U , lower and upper approximations can be introduced in fuzzy set theory to obtain an extended notion called rough fuzzy sets. Alternatively, a fuzzy similarity relation can be used to replace an equivalence relation, resulting in a deviation of rough set theory called fuzzy rough sets.

2.5.4. Applications of rough set

As explained before, Knowledge Discovery in Databases, KDD is the process of extracting the useful information or knowledge from dataset. The process of KDD consists of a few steps where it involves a few number of different tools, methods and underlying theories used in real-world applications. One of them is rough set theory. Its peculiarity is a well understood formal model, which allows finding several kinds of information, such as relevant features or classification rules, using minimal model assumptions. In rough set theory, data model information is stored in a table, where each row or tuple represents a fact or an object. This table is also known as “information

system” and if some of the attributes are interpreted as outcomes of classification, it is called a “decision system”. The data in the real world object is also being stored in such way. However, a huge quantity of data is stored in the table, which is hard to manage from computational point of view. Also it is possible for some facts to not to be consistent to each other. One of the main objectives of rough set data analysis is to reduce the data size. Rough set theory is used to approximate inconsistent information and to exclude redundant data.

The following show the processes in KDD where rough set theory is applied:

- i. Data cleaning and preprocessing:
Data cleaning and preprocessing is one of the application fields of rough sets, it is known that prior knowledge is very important in the KDD process. In KDD, rough sets are used to reduce and clean data with minimal model assumptions. Then the result is used as a basis for further analysis performed with other methods. Also, rough set theory is useful to handle data reduction; missing value; feature selection; and feature extraction.
- ii. Data mining:
Data mining includes many different tasks. Below are three of them that apply rough set theory:
- iii. Classification:
Classification is the original goal of rough set theory. This involves assigning data into classes known as priori.
- iv. Clustering:
Clustering involves grouping together data without the priori. The problem is that overlapping classes may exist in the real world. Due to this data coarseness, rough set theory may be used.
- v. Association rules:
Rough sets attribute dependency analysis may be used to quantify numerically association rules, on the basis of information in data.

2.6. Rough Clustering

This section presents rough clustering and rough categorical data clustering.

2.6.1. Application of rough set in data clustering

As mentioned earlier, rough set theory is one of the method and theory used in the steps involved in KDD process. Among all the steps involve with rough sets, one of them is the process of clustering in the data mining step. A rough cluster is defined in a similar manner to a rough set. The lower approximation of a rough cluster contains objects that only belong to that cluster while the upper approximation of a rough cluster contains objects in the cluster which are also members of the other clusters. The advantage of using rough sets is that, unlike any other techniques, rough set theory does not require any prior information about the data such as a priori probability in statistics and a membership function in fuzzy set theory. This enables the clustering process which involves finding the relationship or associations of data while at the same time having the class of the data undefined.

2.6.2. Rough set theory in categorical data clustering

The problem with categorical data clustering is that the categorical data have multi-valued attributes. This is different when dealing with numerical data clustering as they have attributes with numerical domains which are very easy to handle and very easy to define similarity of them. Unlike the similarity of the numerical data, the similarity of the categorical data is defined as the common objects, common values for the attributes or the association between two. This is where rough set theory is applied. Rough sets are used to develop efficient heuristics searching for relevant tolerance relations that allow extracting interesting patterns of data. An attribute-oriented rough sets technique reduces computational complexity of learning processes and eliminates the unimportant or irrelevant attributes so that the knowledge discovery in database can be efficiently learned. Using rough sets has been shown to be effective for revealing relationships within imprecise data, discovering dependencies among objects and attributes, evaluating the classificatory importance of attributes, removing data re-abundances, and generating decision rules.

Presently, many clustering algorithms or techniques have been proposed but the implementation of them are challenging when dealing with categorical data. Some of the algorithms available at present cannot handle categorical data and some others unable to handle the uncertainty. Many of them have the stability problem and also have efficiency issues. Therefore, new algorithms are made to cluster categorical data which also deals with the uncertainty. Before, the only algorithms which aimed at handling uncertainty in clustering process were based on fuzzy set theory. In 2007, an algorithm, termed MMR was proposed, which uses the rough set theory concepts to deal with the problems. Later in 2009, this algorithm was further improved to develop the algorithm MMeR and it could handle hybrid data. Again, very recently in 2011 MMeR is again improved to develop an algorithm called SDR, which also can handle hybrid data. Both MMeR and SDR can handle both uncertainties and deal with categorical data at the same time but SDR has more efficiency over MMeR and MMR.

CHAPTER 3

METHODOLOGY

This chapter briefly discusses about the model and method of data clustering based on Rough Set Theory (RST). The method so-called Min-Min Roughness is presented details here. There are three main sections in this chapter. The first section is rough set theory, which describes about information system, indiscernibility relations, approximation space and set approximations. The next section describes min-min roughness (MMR) technique, together with an example of the application of the technique on diabetic dataset. The last section describes the object splitting model.

3.1. Rough Set Theory

The problem of imprecise knowledge has been tackled for a long time by mathematicians. Recently it became a crucial issue for computer scientists, particularly in the area of artificial intelligence. There are many approaches to the problem of how to understand and manipulate imprecise knowledge. The most successful one is, no doubt, the fuzzy set theory proposed by Zadeh (Zadeh, 1965). The basic tools of the theory are possibility measures. There is extensive literature on fuzzy logic with also discusses

some of the problem with this theory. The basic problem of fuzzy set theory is the determination of the grade of membership of the value of possibility (Busse, 1998). In the 1980's, Pawlak introduced rough set theory to deal this problem (Pawlak, 1982). Similarly to rough set theory it is not an alternative to classical set theory but it is embedded in it. Fuzzy and rough sets theories are not competitive, but complementary to each other (Pawlak and Skowron, 2007; Pawlak, 1985). Rough set theory has attracted attention to many researchers and practitioners all over the world, who contributed essentially to its development and applications. The original goal of the rough set theory is induction of approximations of concepts. The idea consists of approximation of a subset by a pair of two precise concepts called the *lower approximation* and *upper approximation*. Intuitively, the lower approximation of a set consists of all elements that surely belong to the set, whereas the upper approximation of the set constitutes of all elements that possibly belong to the set. The difference of the upper approximation and the lower approximation is a *boundary region*. It consists of all elements that cannot be classified uniquely to the set or its complement, by employing available knowledge. Thus any rough set, in contrast to a crisp set, has a non-empty boundary region. Motivation for rough set theory has come from the need to represent a subset of a universe in terms of equivalence classes of a partition of the universe. In this chapter, the basic concept of rough set theory in terms of data is presented.

3.1.1. Information System

Data are often presented as a table, columns of which are labeled by *attributes*, rows by *objects* of interest and entries of the table are *attribute values*. By an *information system*, we mean a 4-tuple (quadruple) $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of attribute a , $f : U \times A \rightarrow V$ is a total function such that $f(u, a) \in V_a$, for every $(u, a) \in U \times A$, called information (knowledge) function. An information system is also called a knowledge representation systems or an attribute-valued system and can be intuitively expressed in terms of an information table (refer to Table 3.1).

U	a_1	a_2	\dots	a_k	\dots	$a_{ A }$
u_1	$f(u_1, a_1)$	$f(u_1, a_2)$	\dots	$f(u_1, a_k)$	\dots	$f(u_1, a_{ A })$
u_2	$f(u_2, a_1)$	$f(u_2, a_2)$	\dots	$f(u_2, a_k)$	\dots	$f(u_2, a_{ A })$
u_3	$f(u_3, a_1)$	$f(u_3, a_2)$	\dots	$f(u_3, a_k)$	\dots	$f(u_3, a_{ A })$
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
$u_{ U }$	$f(u_{ U }, a_1)$	$f(u_{ U }, a_2)$	\dots	$f(u_{ U }, a_k)$	\dots	$f(u_{ U }, a_{ A })$

Table 3.1: An information system

In many applications, there is an outcome of classification that is known. This *a posteriori* knowledge is expressed by one (or more) distinguished attribute called decision attribute; the process is known as supervised learning. An information system of this kind is called a decision system. A *decision system* is an information system of the form $D = (U, A = C \cup D, V, f)$, where D is the set of *decision attributes* and $C \cap D = \emptyset$. The elements of C are called *condition attributes*. A simple example of decision system is given in Table 3.2.

Example 3.1.

Suppose that data about 10 diabetics is given, as shown in Table 2.2.

Patient	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Diabetes Type
P1	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	Type 1
P2	No	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Type 1
P3	No	Yes	Yes	Yes	No	No	Yes	No	Yes	No	Type 1
P4	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	Type 1
P5	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Type 1
P6	Yes	Yes	No	Yes	No	No	Yes	Yes	No	No	Type 1
P7	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Type 2
P8	No	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Type 2
P9	No	Yes	No	No	No	Yes	Yes	Yes	No	Yes	Type 2
P10	Yes	No	Yes	No	No	Yes	No	No	Yes	Yes	Type 2

Table 3.2: A diabetic decision system

Where

S1 = Often Thirst

S2 = Excessive Hunger

S3 = Frequent Urination

S4 = Tiredness and Fatigue

S5 = Rapid and/or Sudden Weight Loss

S6 = Blurred Vision

S7 = Numbness and/or Tingling In the Hands and Feet

S8 = Slow healing of Minor-to-treat Yeast Infection in Women

S9 = Recurrent or Hard-o-treat Yeast Infection in Women

S10 = Dry or Itchy Skin

The following values are obtained from Table 3.2,

$$U = \{P1, P2, P3, P4, P5, P6, P7, P8, P9, P10\},$$

$$A = \{S1, S2, S3, S4, S5, S6, S7, S, S9, \text{Diabetes type}\}$$

, where

$$C = \{S1, S2, S3, S4, S5, S6, S7, S8, S9, S10\}, D = \{\text{Diabetes type}\},$$

$$V_{S1} = \{\text{Yes, No}\},$$

$$V_{S2} = \{\text{Yes, No}\},$$

$$V_{S3} = \{\text{Yes, No}\},$$

$$V_{S4} = \{\text{Yes, No}\},$$

$$V_{S5} = \{\text{Yes, No}\},$$

$$V_{S6} = \{\text{Yes, No}\},$$

$$V_{S7} = \{\text{Yes, No}\},$$

$$V_{S8} = \{\text{Yes, No}\},$$

$$V_{S9} = \{\text{Yes, No}\},$$

$$V_{S10} = \{\text{Yes, No}\},$$

$$V_{\text{Diabetes type}} = \{\text{Type 1, Type 2}\}.$$

A relational database may be considered as an information system in which rows are labeled by the objects (entities), columns are labeled by attributes and the entry in row u and column a has the value $f(u, a)$. It is noted that each map

$$f(u, a): U \times A \rightarrow V \text{ is a tuple } t_i = (f(u_i, a_1), f(u_i, a_2), f(u_i, a_3), \dots, f(u_i, a_{|A|})), \text{ for}$$

$1 \leq i \leq |U|$, where $|X|$ is the cardinality of X . Note that the tuple t is not necessarily

associated with entity uniquely. In an information table, two distinct entities could have the same tuple representation (duplicated/redundant tuple), which is *not permissible* in

relational databases. Thus, the concepts in information systems are a generalization of the same concepts in relational databases.

3.1.2. Indiscernibility Relation

From Table 3.2, it is noted that diabetics P1, P2 and P3 are indiscernible (or similar or indistinguishable) with respect to the attribute S1, S2, S3, S6, S7, S8 and Diabetes type. Meanwhile, diabetics P1, P2, P3, P7, P8, P9 are indiscernible with respect to attributes S1 and diabetics P2, P6 and P9 are indiscernible with respect to attributes S2, S7 and S9. Also there are more indiscernibilities that can be seen from the table.

The starting point of rough set theory is the indiscernibility relation, which is generated by information about objects of interest. The indiscernibility relation is intended to express the fact that due to the lack of knowledge we are unable to discern some objects employing the available information. Therefore, generally, we are unable to deal with single object. Nevertheless, we have to consider clusters of indiscernible objects. The following definition precisely defines the notion of indiscernibility relation between two objects.

Definition 2.1. Let $S = (U, A, V, f)$ be an information system and let B be any subset of A . Two elements $x, y \in U$ are said to be B -indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x, a) = f(y, a)$, for every $a \in B$.

Obviously, every subset of A induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute B , denoted by $IND(B)$, is an equivalence relation. It is well known that, an equivalence relation induces unique partition. The partition of U induced by $IND(B)$ in $S = (U, A, V, f)$ denoted by U/B and the equivalence class in the partition U/B containing $x \in U$, denoted by $[x]_B$. Studies of rough set theory may be divided into two class, representing the set-oriented (constructive) and operator-oriented (descriptive) views. They produce extension of

crisp set theory (Yao, 1996; Yao, 1998; Yao, 2001). In this work, rough set theory is presented from the point of view of a constructive approach.

3.1.3. Approximation Space

Let $S = (U, A, V, f)$ be an information system, let B be any subset of A and $IND(B)$ is an indiscernibility relation generated by B on U .

Definition 2.2. An ordered pair $AS = (U, IND(B))$ is called a (Pawlak) approximation space.

Let $x \in U$, the equivalence class of U containing x with respect to R is denoted by $[x]_B$. The family of definable sets, i.e. finite union of arbitrary equivalence classes in partition $U / IND(B)$ in AS , denoted by $DEF(AS)$ is a Boolean algebra (Pawlak, 1982). Thus, an approximation space defines unique topological space, called a *quasi-discrete (clopen) topological space* (Herawan and Mat Deris, 2009a). Given arbitrary subset $X \subseteq U$, X may not be presented as union of some equivalence classes in U . In other means that a subset X cannot be described precisely in AS . Thus, a subset X may be characterized by a pair of its approximations, called lower and upper approximations. It is here that the notion of rough set emerges.

3.1.4. Set Approximations

The indiscernibility relation will be used to define set approximations that are the basic concepts of rough set theory. The notions of lower and upper approximations of a set can be defined as follows.

Definition 2.3. Let $S = (U, A, V, f)$ be an information system, let B be any subset of A and let X be any subset of U . The B -lower approximation of X , denoted by $\underline{B}(X)$ and B -upper approximations of X , denoted by $\overline{B}(X)$, respectively, are defined by

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\}.$$

From Definition 2.3, the following interpretations are obtained

- a. The *lower approximation* of a set X with respect to B is the set of all objects, which can be for *certain* classified as X using B (are certainly X in view of B).
- b. The *upper approximation* of a set X with respect to B is the set of all objects which can be *possibly* classified as X using B (are possibly X in view of B).

Hence, with respect to arbitrary subset $X \subseteq U$, the universe U can be divided into three disjoint regions using the lower and upper approximations

- a. The *positive region* $\text{POS}_B(X) = \underline{B}(X)$, i.e., the set of all objects, which can be for *certain* classified as X using B (are *certainly* X with respect to B).
- b. The *boundary region* $\text{BND}_B(X) = \overline{B}(X) - \underline{B}(X)$, i.e., the set of all objects, which can be classified neither as X nor as not- X using B .
- c. The *negative region* $\text{NEG}_B(X) = U - \overline{B}(X)$, i.e., the set of all objects, which can be for *certain* classified as not- X using B (are *certainly* not- X with respect to B).

These notions of lower and upper approximations can be shown clearly as in Figure 3.1.

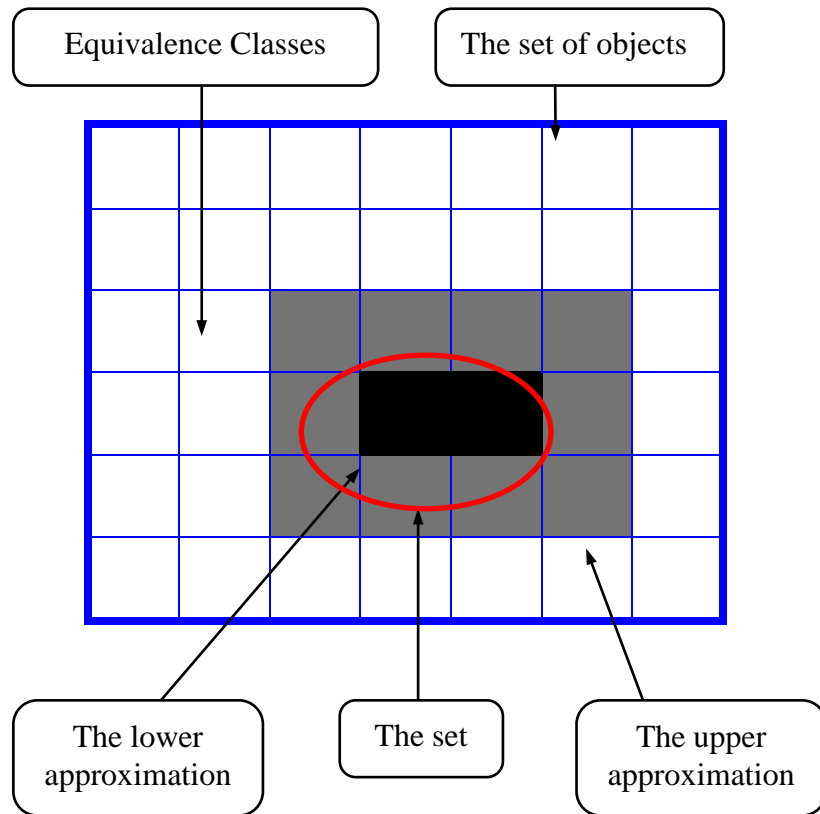


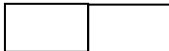


Figure 3.1: Set approximations

From Figure 3.1, three disjoint regions are given as follows

- a. The positive region 
- b. The boundary region 
- c. The negative region 

It is easily seen that the lower and the upper approximations of a set, respectively, are *interior* and *closure* operations in a quasi discrete topology generated by the indiscernibility relation [Herawan and Mat Deris, 2009e].

The accuracy of approximation (accuracy of roughness) of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted $\alpha_B(X)$ is measured by

$$\alpha_B(X) = \frac{|B(X)|}{|\overline{B}(X)|}, \quad (2.1)$$

where $|X|$ denotes the cardinality of X . For empty set ϕ , it is defined that $\alpha_B(\phi) = 1$ (Pawlak and Skowron, 2007). Obviously, $0 \leq \alpha_B(X) \leq 1$. If X is a union of some equivalence classes of U , then $\alpha_B(X) = 1$. Thus, the set X is *crisp* (precise) with respect to B . And, if X is not a union of some equivalence classes of U , then $\alpha_B(X) < 1$. Thus, the set X is *rough* (imprecise) with respect to B (Pawlak and Skowron, 2007). This means that the higher of accuracy of approximation of any subset $X \subseteq U$ is the more precise (the less imprecise) of itself.

Example 2.2. Let us depict above notions by examples referring to Table 3.2. Consider the concept “Decision”, i.e., the set $X(\text{Diabetes type} = \text{Type 1}) = \{P1, P2, P3, P4, P5, P6\}$ and the set of attributes $C = \{S2, S7, S9\}$. The partition of U induced by $IND(C)$ is given by

$$U / C = \{\{P1, P3\}, \{P2, P6, P9\}, \{P4\}, \{P5, P8\}, \{P7\}, \{P10\}\}.$$

The corresponding lower approximation and upper approximation of X are as follows

$$\underline{C}(X) = \{P1, P3, P4\} \text{ and } \overline{C}(X) = \{P1, P2, P3, P4, P5, P6, P8, P9\}.$$

Thus, concept “Decision” is imprecise (rough). For this case, the accuracy of approximation is given as

$$\alpha_c(X) = \frac{3}{8}.$$

It means that the concept “Decision” can be characterized partially employing attributes Analysis, Algebra and Statistics.

The accuracy of roughness in Equation (2.1) can also be interpreted using the well-known Marczewski-Steinhaus (MZ) metric (Yao, 1996; Yao, 1998; Yao, 2001). Let $S = (U, A, V, f)$ be an information system and given two subsets $X, Y \subseteq U$, the MZ metric measuring the distance X and Y is defined as

$$D(X, Y) = \frac{|X \Delta Y|}{|X \cup Y|},$$

where, $X \Delta Y = (X \cup Y) - (X \cap Y)$ denotes the symmetric difference between two sets X and Y .

Therefore, the MZ metric can be expressed as

$$\begin{aligned} D(X, Y) &= \frac{(X \cup Y) - (X \cap Y)}{|X \cup Y|} \\ &= 1 - \frac{|X \cap Y|}{|X \cup Y|}. \end{aligned}$$

Notice that,

- a. If X and Y are totally different, i.e. $X \cap Y = \phi$ (in other words X and Y are disjoint), then the metric reaches the maximum value of 1
- b. If X and Y are exactly the same, i.e. $X = Y$, then the metric reaches minimum value of 0.

By applying the MZ metric to the lower and upper approximations of a subset $X \subseteq U$ in information system S , the following MZ metric is obtained

$$\begin{aligned}
 D(\underline{B}(X), \overline{B}(X)) &= 1 - \frac{|\underline{B}(X) \cap \overline{B}(X)|}{|\underline{B}(X) \cup \overline{B}(X)|}, \\
 &= 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|}, \\
 &= 1 - \alpha_B(X). \tag{2.2}
 \end{aligned}$$

The accuracy of roughness may be viewed as an inverse of MZ metric when applied to lower and upper approximations. In other words, the distance between the lower and upper approximations determines the accuracy of the rough set approximations.

3.2. Min-Min Roughness

In this section, a technique for selecting a clustering attribute based on rough set theory is presented. There are a few techniques that had been proposed to deal with the clustering attribute selection. Mazlack et al. proposed two techniques to select clustering attribute, which is bi-clustering (BC) technique and total roughness (TR) technique. Then, Parmar et al. proposes a technique called min-min roughness (MMR) which improves the BC technique for data set with multi-valued attributes. Another techniques also had been proposed, which is called maximum dependency of attributes (MDA). This section, however, will be presenting the technique MMR to select the clustering attributes.

3.2.1. Selecting a clustering attribute

To find meaningful clusters from a dataset, clustering attribute is conducted so that attributes within the clusters made will have a high correlation or high interdependence

to each other while the attributes in other clusters are less correlated or more independent.

3.2.2. Model for selecting a clustering attribute

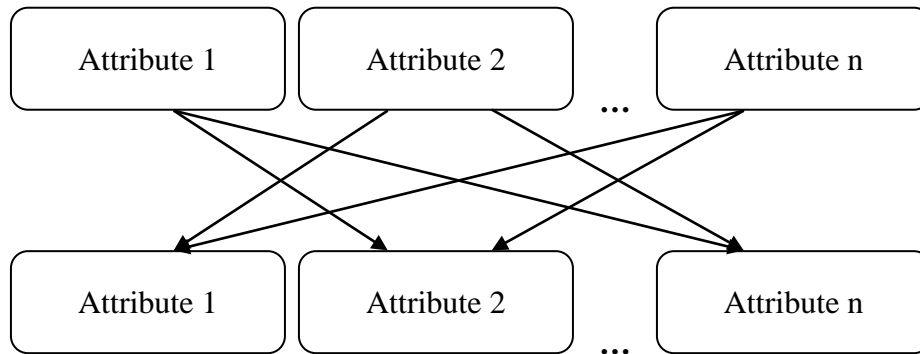


Figure 3.2: Model for selecting a clustering attribute

3.2.3. Min-Min Roughness Technique

The following Table shows step-by-step to calculate Min-Min Roughness.

Step	Min-Min Roughness
1	Given data set
2	Each attribute in data set considered as a candidate attribute to partition
3	Determine equivalence classes of attribute-value pairs
4	Determine lower approximation of each equivalence classes in attribute a_i w.r.t. to attribute a_j , $i \neq j$
5	Determine upper approximation of each equivalence classes in attribute a_i w.r.t. to attribute a_j , $i \neq j$
6	Calculate roughness of each equivalence classes in attribute a_i w.r.t. to attribute a_j , $i \neq j$
7	Calculate mean roughness of attribute a_i w.r.t. to attribute a_j , $i \neq j$
8	Calculate minimum roughness a_i w.r.t. to all attribute a_j , $i \neq j$
9	If there are two greatest value of mean roughness, calculate

	minimum roughness relative to the second, third greater minimum roughness until the tie is broken
10	Selecting a clustering attribute

Table 3.3: Step-by-step Max-Max Roughness

3.2.4. Algorithm

Below show the example algorithm to obtain the MMR from the diabetics information dataset.

Input: Diabetic information dataset

// finding the U/IND for each attribute

```

x=1
for i = 1 to nth attribute
  set att(i,1) as set(i,x)
  for j = 1 to nth row
    for k = 1 to nth
      if att(i,j) doesn't belong to any set & j ≠ k
        then if att(i,j) = att(i,k) & att(i,k) belong to a set
              then set att(i,j) as same set as att(i,k)
              else set att(i,j) as set(i,x++)
        end if
      end if
    end for loop
  end for loop
end for loop

```

// finding the number of element in lower and upper approximation for each attribute

```

for i = 1 to nth attribute
  for j = 1 to nth attribute
    for k = 1 to nth attributeSet
      if set(j,k) ∈ set(i,k) & i ≠ j
        then lowerApprox(ai,k) =
              lowerApprox(ai,k) + no of element in
              set(j,k)
      else if set(i,k) ∈ & ij
        then upperApprox(ai,k) =
              upperApprox(ai,k) + no of element in
              set(j,k)
      end if
    end for loop
  end for loop
end for loop

```

```

end for loop

// calculating roughness for each attribute set
for i = 1 to nth attribute
    for k = 1 to nth attributeSet
        roughness(ai,k) =
            1 - (lowerApprox(ai,k) ÷ upperApprox(ai,k))
    end for loop
end for loop

// calculating mean roughness for each attribute
x=1
for i = 1 to nth attribute
    for k = 1 to nth attributeSet
        totalRoughness() =
            totalRoughness(ai) + roughness(ai,k)
        totalAttribute(ai) = totalAttribute(ai) + x
    end for loop
end for loop

for i = 1 to nth attribute
    meanRoughness(ai) = totalRoughness(ai) ÷ totalAttribute(ai)
end for loop

// finding the min roughness from all attributes
for i = 2 to nth attribute
    if meanRoughness(ai) < meanRoughness(ai-1)
        then minRoughness = meanRoughness(ai)
    end if
end for loop

// for case where the lowest minimum meanRoughness is more than 1
count = 0
for i = 1 to nth attribute
    if minRoughness = meanRoughness(ai)
        then count = count + 1
    end if
end for loop

if count > 1
    then for i = 2 to nth attribute
        if meanRoughness(ai) < meanRoughness(ai-1)
            & meanRoughness(ai) ≠ minRoughness
        then min-minRoughness =
            meanRoughness(ai)
        end if
    end for loop

```

```
count = 0
for  $i = 1$  to  $n$ th attribute
  if min-minRoughness = meanRoughness( $a_i$ )
  then count = count + 1
  end if
end for loop
end if
```

3.2.5. Example

The following example shows a calculation result of Max-Max Roughness through an information system for diabetic decision system.

U	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
1	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No
2	No	Yes	Yes	No	Yes	No	Yes	No	No	Yes
3	No	Yes	Yes	Yes	No	No	Yes	No	Yes	No
4	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes
5	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No
6	Yes	Yes	No	Yes	No	No	Yes	Yes	No	No
7	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
8	No	Yes	No	No	No	Yes	No	Yes	Yes	Yes
9	No	Yes	No	No	No	Yes	Yes	Yes	No	Yes
10	Yes	No	Yes	No	No	Yes	No	No	Yes	Yes

Table 3.4: An information system in MMR.

As an information system, from Table 3.4, we have:

$$U = \{1,2,3,4,5,6,7,8,9,10\},$$

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}\},$$

$$Va_1 = \{\text{yes, no}\},$$

$$Va_2 = \{\text{yes, no}\},$$

$$Va_3 = \{\text{yes, no}\},$$

$$Va_4 = \{\text{yes, no}\},$$

$$Va_5 = \{\text{yes, no}\},$$

$$Va_6 = \{\text{yes, no}\},$$

$$Va_7 = \{\text{yes, no}\},$$

$$Va_8 = \{\text{yes, no}\},$$

$$Va_9 = \{\text{yes, no}\},$$

$$Va_{10} = \{\text{yes, no}\}.$$

Seven partitions of U generated by indiscernibility relation of singleton attribute are;

- a. $X(a_1 = \text{yes}) = \{4,5,6,10\}$, $X(a_1 = \text{no}) = \{1,2,3,7,8,9\}$,
 $U / IND(a_1) = \{\{4,5,6,10\}, \{1,2,3,7,8,9\}\}.$
- b. $X(a_2 = \text{yes}) = \{1,2,3,5,6,8,9\}$, $X(a_2 = \text{no}) = \{4,7,10\}$,
 $U / IND(a_2) = \{\{1,2,3,5,6,8,9\}, \{4,7,10\}\}.$
- c. $X(a_3 = \text{yes}) = \{1,2,3,4,10\}$, $X(a_3 = \text{no}) = \{5,6,7,8,9\}$,
 $U / IND(a_3) = \{\{1,2,3,4,10\}, \{5,6,7,8,9\}\}.$
- d. $X(a_4 = \text{yes}) = \{3,5,6\}$, $X(a_4 = \text{no}) = \{1,2,4,7,8,9,10\}$,
 $U / IND(a_4) = \{\{3,5,6\}, \{1,2,4,7,8,9,10\}\}.$
- e. $X(a_5 = \text{yes}) = \{1,2,4\}$, $X(a_5 = \text{no}) = \{3,5,6,7,8,9,10\}$,
 $U / IND(a_5) = \{\{1,2,4\}, \{3,5,6,7,8,9,10\}\}.$
- f. $X(a_6 = \text{yes}) = \{5,7,8,9,10\}$, $X(a_6 = \text{no}) = \{1,2,3,4,6\}$,
 $U / IND(a_6) = \{\{5,7,8,9,10\}, \{1,2,3,4,6\}\}.$
- g. $X(a_7 = \text{yes}) = \{1,2,3,6,7,9\}$, $X(a_7 = \text{no}) = \{4,5,8,10\}$,
 $U / IND(a_7) = \{\{1,2,3,6,7,9\}, \{4,5,8,10\}\}.$
- h. $X(a_8 = \text{yes}) = \{4,6,7,8,9\}$, $X(a_8 = \text{no}) = \{1,2,3,5,10\}$,
 $U / IND(a_8) = \{\{4,6,7,8,9\}, \{1,2,3,5,10\}\}.$
- i. $X(a_9 = \text{yes}) = \{1,3,5,7,8,10\}$, $X(a_9 = \text{no}) = \{2,4,6,9\}$,
 $U / IND(a_9) = \{\{1,3,5,7,8,10\}, \{2,4,6,9\}\}.$
- j. $X(a_{10} = \text{yes}) = \{2,4,7,8,9,10\}$, $X(a_{10} = \text{no}) = \{1,3,5,6\}$,
 $U / IND(a_{10}) = \{\{2,4,7,8,9,10\}, \{1,3,5,6\}\}.$

Calculation of roughness and mean roughness

First, we determine of upper and lower approximations of singleton attribute with respect to other different singleton attribute. Then we calculate the roughness and the mean roughness of each attribute.

a. **Attribute a_1**

For attribute a_1 , it is clear that $|V(a_1)| = 2$. The roughness and the mean roughness on a_1 with respect to a_i , $i = 2,3,4,5,6,7,8,9,10$ is calculated as the following.

1) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_2(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_1) = \frac{1+1}{2} = 1.$$

2) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_1) = \frac{1+1}{2} = 1.$$

3) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_4(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_4}(a_1) = \frac{1+1}{3} = 1.$$

4) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_5(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_5}(a_1) = \frac{1+1}{3} = 1.$$

5) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_1) = \frac{1+1}{3} = 1.$$

6) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_1) = \frac{1+1}{3} = 1.$$

7) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_1) = \frac{1+1}{3} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_1) = \frac{1+1}{3} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_1 = yes) = \phi \text{ and } \overline{X}(a_1 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_1 = no) = \phi \text{ and } \overline{X}(a_1 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_1 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_1 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_1) = \frac{1+1}{3} = 1.$$

b. **Attribute a_2**

For attribute a_2 , it is clear that $|V(a_2)| = 2$. The roughness and the mean roughness on a_2 with respect to a_i , $i = 1, 3, 4, 5, 6, 7, 8, 9, 10$ is calculated as the following

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_2 = yes) = \phi \text{ and } \overline{X}(a_2 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_2 = no) = \phi \text{ and } \overline{X}(a_2 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_1(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_2) = \frac{1+1}{2} = 1.$$

2) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_2 = yes) = \phi \text{ and } \overline{X}(a_2 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_2 = no) = \phi \text{ and } \overline{X}(a_2 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_3(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_2) = \frac{1+1}{2} = 1.$$

3) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_2 = yes) = \{3, 5, 6\} \text{ and } \overline{X}(a_2 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_2 = no) = \phi \text{ and } \overline{X}(a_2 = no) = \{1,2,4,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_2 = yes) = 1 - \frac{3}{10} = 0.7,$$

$$Ra_4(X|a_2 = no) = 1 - \frac{0}{7} = 1.$$

Mean roughness

$$Rough_{a_4}(a_2) = \frac{0.7+1}{2} = 0.85.$$

4) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_2 = yes) = \phi \text{ and } \overline{X}(a_2 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_2 = no) = \phi \text{ and } \overline{X}(a_2 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_5(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_5}(a_2) = \frac{1+1}{2} = 1.$$

5) With respect to a_6

$$\underline{X}(a_2 = yes) = \phi \text{ and } \overline{X}(a_2 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_2 = no) = \phi \text{ and } \overline{X}(a_2 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_2) = \frac{1+1}{2} = 1.$$

6) With respect to a_7

$$\underline{X(a_2 = yes)} = \phi \text{ and } \overline{X(a_2 = yes)} = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X(a_2 = no)} = \phi \text{ and } \overline{X(a_2 = no)} = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_2) = \frac{1+1}{2} = 1.$$

7) With respect to a_8

$$\underline{X(a_2 = yes)} = \phi \text{ and } \overline{X(a_2 = yes)} = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X(a_2 = no)} = \phi \text{ and } \overline{X(a_2 = no)} = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_2) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

$$\underline{X(a_2 = yes)} = \phi \text{ and } \overline{X(a_2 = yes)} = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X(a_2 = no)} = \phi \text{ and } \overline{X(a_2 = no)} = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_2 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_2 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_2) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

$$\underline{X}(a_2 = yes) = \{1,3,5,6\} \text{ and } \overline{X}(a_2 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_2 = no) = \phi \text{ and } \overline{X}(a_2 = no) = \{2,4,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_2 = yes) = 1 - \frac{4}{10} = 0.6,$$

$$Ra_{10}(X|a_2 = no) = 1 - \frac{0}{6} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_2) = \frac{0.6+1}{2} = 0.8.$$

c. **Attribute a_3**

For attribute a_3 , it is clear that $|V(a_3)| = 2$. The roughness and the mean roughness on a_3 with respect to a_i , $i = 1,2,4,5,6,7,8,9,10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_1(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_3) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_2(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_3) = \frac{1+1}{2} = 1.$$

3) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_4(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_4}(a_3) = \frac{1+1}{2} = 1.$$

4) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \{1,2,4\} \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{3,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_3 = yes) = 1 - \frac{3}{10} = 0.7,$$

$$Ra_5(X|a_3 = no) = 1 - \frac{0}{7} = 1.$$

Mean roughness

$$Rough_{a_5}(a_3) = \frac{0.7+1}{2} = 0.85.$$

5) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_3) = \frac{1+1}{2} = 1.$$

6) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_3) = \frac{1+1}{2} = 1.$$

7) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_3) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_3) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_3 = yes) = \phi \text{ and } \overline{X}(a_3 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_3 = no) = \phi \text{ and } \overline{X}(a_3 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_3 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_3 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_3) = \frac{1+1}{2} = 1.$$

d. **Attribute a_4**

For attribute a_4 , it is clear that $|V(a_4)| = 2$. The roughness and the mean roughness on a_4 with respect to a_i , $i = 1, 2, 3, 5, 6, 7, 8, 9, 10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_4 = no) = \phi \text{ and } \overline{X}(a_4 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_1(X|a_4 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_4 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_4) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1, 2, 3, 5, 6, 8, 9\},$$

$$\underline{X}(a_4 = no) = \{4, 7, 10\} \text{ and } \overline{X}(a_4 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_2(X|a_4 = \text{yes}) = 1 - \frac{0}{7} = 1,$$

$$Ra_2(X|a_4 = \text{no}) = 1 - \frac{3}{10} = 0.7.$$

Mean roughness

$$Rough_{a_2}(a_4) = \frac{1+0.7}{2} = 0.85.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_4 = \text{yes}) = \phi \text{ and } \overline{X}(a_4 = \text{yes}) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_4 = \text{no}) = \phi \text{ and } \overline{X}(a_4 = \text{no}) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_4 = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_4 = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_4) = \frac{1+1}{2} = 1.$$

4) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_4 = \text{yes}) = \phi \text{ and } \overline{X}(a_4 = \text{yes}) = \{3,5,6,7,8,9,10\},$$

$$\underline{X}(a_4 = \text{no}) = \{1,2,4\} \text{ and } \overline{X}(a_4 = \text{no}) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_4 = \text{yes}) = 1 - \frac{0}{7} = 1,$$

$$Ra_5(X|a_4 = \text{no}) = 1 - \frac{3}{10} = 0.7.$$

Mean roughness

$$Rough_{a_5}(a_4) = \frac{1+0.7}{2} = 0.85.$$

5) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_4 = no) = \phi \text{ and } \overline{X}(a_4 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_4 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_4 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_4) = \frac{1+1}{2} = 1.$$

6) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_4 = no) = \phi \text{ and } \overline{X}(a_4 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_4 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_4 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_4) = \frac{1+1}{2} = 1.$$

7) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_4 = no) = \phi \text{ and } \overline{X}(a_4 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_4 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_4 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_4) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_4 = no) = \phi \text{ and } \overline{X}(a_4 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_4 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_4 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_4) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_4 = yes) = \phi \text{ and } \overline{X}(a_4 = yes) = \{1,3,5,6\},$$

$$\underline{X}(a_4 = no) = \{2,4,7,8,9,10\} \text{ and } \overline{X}(a_4 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_4 = yes) = 1 - \frac{0}{4} = 1,$$

$$Ra_{10}(X|a_4 = no) = 1 - \frac{6}{10} = 0.4.$$

Mean roughness

$$Rough_{a_{10}}(a_4) = \frac{1+0.4}{2} = 0.7.$$

e. **Attribute a_5**

For attribute a_5 , it is clear that $|V(a_5)| = 2$. The roughness and the mean roughness on a_5 with respect to a_i , $i = 1, 2, 3, 4, 6, 7, 8, 9, 10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_5 = no) = \phi \text{ and } \overline{X}(a_5 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_1(X|a_5 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_5 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_5) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_5 = no) = \phi \text{ and } \overline{X}(a_5 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_2(X|a_5 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_5 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_5) = \frac{1+1}{2} = 1.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,3,4,10\},$$

$$\underline{X}(a_5 = no) = \{5,6,7,8,9\} \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_5 = yes) = 1 - \frac{0}{5} = 1,$$

$$Ra_3(X|a_5 = no) = 1 - \frac{5}{10} = 0.5.$$

Mean roughness

$$Rough_{a_3}(a_5) = \frac{1+0.5}{2} = 0.75.$$

4) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,4,7,8,9,10\},$$

$$\underline{X}(a_5 = no) = \{3,5,6\} \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_5 = yes) = 1 - \frac{0}{7} = 1,$$

$$Ra_4(X|a_5 = no) = 1 - \frac{3}{10} = 0.7.$$

Mean roughness

$$Rough_{a_4}(a_5) = \frac{1+0.7}{2} = 0.85.$$

5) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,3,4,6\},$$

$$\underline{X}(a_5 = no) = \{5,7,8,9,10\} \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_5 = yes) = 1 - \frac{0}{5} = 1,$$

$$Ra_6(X|a_5 = no) = 1 - \frac{5}{10} = 0.5.$$

Mean roughness

$$Rough_{a_6}(a_5) = \frac{1+0.5}{2} = 0.75.$$

6) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_5 = no) = \phi \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_5 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_5 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_5) = \frac{1+1}{2} = 1.$$

7) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_5 = no) = \phi \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_5 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_5 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_5) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_5 = no) = \phi \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_5 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_5 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_5) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_5 = yes) = \phi \text{ and } \overline{X}(a_5 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_5 = no) = \phi \text{ and } \overline{X}(a_5 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_5 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_5 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_5) = \frac{1+1}{2} = 1.$$

f. **Attribute a_6**

For attribute a_6 , it is clear that $|V(a_6)| = 2$. The roughness and the mean roughness on a_6 with respect to a_i , $i = 1,2,3,4,5,7,8,9,10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_1(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_6) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_2(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_6) = \frac{1+1}{2} = 1.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_6) = \frac{1+1}{2} = 1.$$

4) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_4(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_4}(a_6) = \frac{1+1}{2} = 1.$$

5) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{3,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \{1,2,4\} \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_6 = yes) = 1 - \frac{0}{7} = 1,$$

$$Ra_5(X|a_6 = no) = 1 - \frac{3}{10} = 0.7.$$

Mean roughness

$$Rough_{a_5}(a_6) = \frac{1+0.7}{2} = 0.85.$$

6) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_6) = \frac{1+1}{2} = 1.$$

7) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_6) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_6) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_6 = yes) = \phi \text{ and } \overline{X}(a_6 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_6 = no) = \phi \text{ and } \overline{X}(a_6 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_6 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_6 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_6) = \frac{1+1}{2} = 1.$$

g. **Attribute** a_7

For attribute a_7 , it is clear that $|V(a_7)| = 2$. The roughness and the mean roughness on a_7 with respect to a_i , $i = 1,2,3,4,5,6,8,9,10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_1(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_7) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_2(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_7) = \frac{1+1}{2} = 1.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_7) = \frac{1+1}{2} = 1.$$

4) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_4(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_4}(a_7) = \frac{1+1}{2} = 1.$$

5) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_5(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_5}(a_7) = \frac{1+1}{2} = 1.$$

6) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_7) = \frac{1+1}{2} = 1.$$

7) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_7) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_7) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_7 = yes) = \phi \text{ and } \overline{X}(a_7 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_7 = no) = \phi \text{ and } \overline{X}(a_7 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_7 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_7 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_7) = \frac{1+1}{2} = 1.$$

h. **Attribute** a_8

For attribute a_8 , it is clear that $|V(a_8)| = 2$. The roughness and the mean roughness on a_8 with respect to a_i , $i = 1,2,3,4,5,6,7,9,10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_1(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_8) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_2(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_8) = \frac{1+1}{2} = 1.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_8) = \frac{1+1}{2} = 1.$$

4) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_4(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_4}(a_8) = \frac{1+1}{2} = 1.$$

5) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_5(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_5}(a_8) = \frac{1+1}{2} = 1.$$

6) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_8) = \frac{1+1}{2} = 1.$$

7) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_8) = \frac{1+1}{2} = 1.$$

8) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_8) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_8 = yes) = \phi \text{ and } \overline{X}(a_8 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_8 = no) = \phi \text{ and } \overline{X}(a_8 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_8 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_8 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_8) = \frac{1+1}{2} = 1.$$

i. **Attribute a_9**

For attribute a_9 , it is clear that $|V(a_9)| = 2$. The roughness and the mean roughness on a_9 with respect to a_i , $i = 1, 2, 3, 4, 5, 6, 7, 8, 10$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_1(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_9) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Roughness

$$Ra_2(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_2(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_2}(a_9) = \frac{1+1}{2} = 1.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_9) = \frac{1+1}{2} = 1.$$

4) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_4(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_4}(a_9) = \frac{1+1}{2} = 1.$$

5) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_5(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_5}(a_9) = \frac{1+1}{2} = 1.$$

6) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_9) = \frac{1+1}{2} = 1.$$

7) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_9) = \frac{1+1}{2} = 1.$$

8) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_9) = \frac{1+1}{2} = 1.$$

9) With respect to a_{10}

The lower and upper approximations are

$$\underline{X}(a_9 = yes) = \phi \text{ and } \overline{X}(a_9 = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_9 = no) = \phi \text{ and } \overline{X}(a_9 = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_{10}(X|a_9 = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_{10}(X|a_9 = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_{10}}(a_9) = \frac{1+1}{2} = 1.$$

j. **Attribute** a_{10}

For attribute a_{10} , it is clear that $|V(a_{10})| = 2$. The roughness and the mean roughness on a_{10} with respect to a_i , $i = 1,2,3,4,5,6,7,8,9$ is calculated as the following.

1) With respect to a_1

The lower and upper approximations are

$$\underline{X}(a_{10} = yes) = \phi \text{ and } \overline{X}(a_{10} = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = no) = \phi \text{ and } \overline{X}(a_{10} = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_1(X|a_{10} = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_1(X|a_{10} = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_1}(a_{10}) = \frac{1+1}{2} = 1.$$

2) With respect to a_2

The lower and upper approximations are

$$\underline{X}(a_{10} = \text{yes}) = \{4,7,10\} \text{ and } \overline{X}(a_{10} = \text{yes}) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = \text{no}) = \phi \text{ and } \overline{X}(a_{10} = \text{no}) = \{1,2,3,5,6,8,9\}.$$

Roughness

$$Ra_2(X|a_{10} = \text{yes}) = 1 - \frac{3}{10} = 0.7,$$

$$Ra_2(X|a_{10} = \text{no}) = 1 - \frac{0}{7} = 1.$$

Mean roughness

$$Rough_{a_2}(a_{10}) = \frac{0.7+1}{2} = 0.85.$$

3) With respect to a_3

The lower and upper approximations are

$$\underline{X}(a_{10} = \text{yes}) = \phi \text{ and } \overline{X}(a_{10} = \text{yes}) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = \text{no}) = \phi \text{ and } \overline{X}(a_{10} = \text{no}) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_3(X|a_{10} = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_3(X|a_{10} = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_3}(a_{10}) = \frac{1+1}{2} = 1.$$

4) With respect to a_4

The lower and upper approximations are

$$\underline{X}(a_{10} = yes) = \phi \text{ and } \overline{X}(a_{10} = yes) = \{1,2,4,7,8,9,10\},$$

$$\underline{X}(a_{10} = no) = \{3,5,6\} \text{ and } \overline{X}(a_{10} = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_4(X|a_{10} = yes) = 1 - \frac{0}{7} = 1,$$

$$Ra_4(X|a_{10} = no) = 1 - \frac{3}{10} = 0.7.$$

Mean roughness

$$Rough_{a_4}(a_{10}) = \frac{1+0.7}{2} = 0.85.$$

5) With respect to a_5

The lower and upper approximations are

$$\underline{X}(a_{10} = yes) = \phi \text{ and } \overline{X}(a_{10} = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = no) = \phi \text{ and } \overline{X}(a_{10} = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_5(X|a_{10} = yes) = 1 - \frac{0}{10} = 1,$$

$$Ra_5(X|a_{10} = no) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_5}(a_{10}) = \frac{1+1}{2} = 1.$$

6) With respect to a_6

The lower and upper approximations are

$$\underline{X}(a_{10} = yes) = \phi \text{ and } \overline{X}(a_{10} = yes) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = no) = \phi \text{ and } \overline{X}(a_{10} = no) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_6(X|a_{10} = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_6(X|a_{10} = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_6}(a_{10}) = \frac{1+1}{2} = 1.$$

7) With respect to a_7

The lower and upper approximations are

$$\underline{X}(a_{10} = \text{yes}) = \phi \text{ and } \overline{X}(a_{10} = \text{yes}) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = \text{no}) = \phi \text{ and } \overline{X}(a_{10} = \text{no}) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_7(X|a_{10} = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_7(X|a_{10} = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_7}(a_{10}) = \frac{1+1}{2} = 1.$$

8) With respect to a_8

The lower and upper approximations are

$$\underline{X}(a_{10} = \text{yes}) = \phi \text{ and } \overline{X}(a_{10} = \text{yes}) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = \text{no}) = \phi \text{ and } \overline{X}(a_{10} = \text{no}) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_8(X|a_{10} = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_8(X|a_{10} = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_8}(a_{10}) = \frac{1+1}{2} = 1.$$

9) With respect to a_9

The lower and upper approximations are

$$\underline{X}(a_{10} = \text{yes}) = \phi \text{ and } \overline{X}(a_{10} = \text{yes}) = \{1,2,3,4,5,6,7,8,9,10\},$$

$$\underline{X}(a_{10} = \text{no}) = \phi \text{ and } \overline{X}(a_{10} = \text{no}) = \{1,2,3,4,5,6,7,8,9,10\}.$$

Roughness

$$Ra_9(X|a_{10} = \text{yes}) = 1 - \frac{0}{10} = 1,$$

$$Ra_9(X|a_{10} = \text{no}) = 1 - \frac{0}{10} = 1.$$

Mean roughness

$$Rough_{a_9}(a_{10}) = \frac{1+1}{2} = 1.$$

Calculation of MMR on each attribute:

Mean roughness calculation for attribute a_1

With respect to	Yes	No	Mean Roughness
a_2	1	1	1
a_3	1	1	1
a_4	1	1	1
a_5	1	1	1
a_6	1	1	1
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	1	1	1

Table 3.5: Mean roughness a_1

With the same way as in 1, we get mean roughness for attribute a_i , $i = 2,3,4,5,6,7,8,9,10$ below.

a_2 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_3	1	1	1
a_4	0.70	1	0.85
a_5	1	1	1
a_6	1	1	1
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	0.60	1	0.80

Table 3.6: Mean roughness a_2

a_3 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	1	1	1
a_4	1	1	1
a_5	0.70	1	0.85
a_6	1	1	1
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	1	1	1

Table 3.7: Mean roughness a_3

a_4 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	1	0.70	0.85
a_3	1	1	1
a_5	1	0.70	0.85
a_6	1	1	1
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	1	0.40	0.70

Table 3.8: Mean roughness a_4

a_5 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	1	1	1
a_3	1	0.50	0.75
a_4	1	0.70	0.85
a_6	1	0.50	0.75
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	1	1	1

Table 3.9: Mean roughness a_5

a_6 with respect to	Yes	No	Mean Roughness
-----------------------	-----	----	----------------

a_1	1	1	1
a_2	1	1	1
a_3	1	1	1
a_4	1	1	1
a_5	1	0.70	0.85
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	1	1	1

Table 3.10: Mean roughness a_6

a_7 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	1	1	1
a_3	1	1	1
a_4	1	1	1
a_5	1	1	1
a_6	1	1	1
a_8	1	1	1
a_9	1	1	1
a_{10}	1	1	1

Table 3.11: Mean roughness a_7

a_8 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	1	1	1

a_3	1	1	1
a_4	1	1	1
a_5	1	1	1
a_6	1	1	1
a_7	1	1	1
a_9	1	1	1
a_{10}	1	1	1

Table 3.12: Mean roughness a_8

a_9 with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	1	1	1
a_3	1	1	1
a_4	1	1	1
a_5	1	1	1
a_6	1	1	1
a_7	1	1	1
a_8	1	1	1
a_{10}	1	1	1

Table 3.13: Mean roughness a_9

a_{10} with respect to	Yes	No	Mean Roughness
a_1	1	1	1
a_2	0.70	1	0.85
a_3	1	1	1
a_4	1	0.70	0.85
a_5	1	1	1
a_6	1	1	1
a_7	1	1	1
a_8	1	1	1
a_9	1	1	1

Table 3.14: Mean roughness a_{10}

From Tables 3.5 until Table 3.14, we get the Min-Min Roughness on each attribute

Attribute	Mean Roughness									Min Roughness
a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	1
	1	1	1	1	1	1	1	1	1	
a_2	a_1	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	0.80
	1	1	0.85	1	1	1	1	1	0.80	
a_3	a_1	a_2	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	0.85
	1	1	1	0.85	1	1	1	1	1	
a_4	a_1	a_2	a_3	a_5	a_6	a_7	a_8	a_9	a_{10}	0.70
	1	0.85	1	0.85	1	1	1	1	0.70	
a_5	a_1	a_2	a_3	a_4	a_6	a_7	a_8	a_9	a_{10}	0.75
	1	1	0.75	0.85	0.75	1	1	1	1	
a_6	a_1	a_2	a_3	a_4	a_5	a_7	a_8	a_9	a_{10}	0.85
	1	1	1	1	0.85	1	1	1	1	

a_7	a_1	a_2	a_3	a_4	a_5	a_6	a_8	a_9	a_{10}	1
	1	1	1	1	1	1	1	1	1	
a_8	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_9	a_{10}	1
	1	1	1	1	1	1	1	1	1	
a_9	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_{10}	1
	1	1	1	1	1	1	1	1	1	
a_{10}	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	0.85
	1	0.85	1	0.85	1	1	1	1	1	

Table 3.15: Minimum roughness

3.3. Object Splitting model

From Table 3.15, we have calculated the Min-Min Roughness of all attributes

	Mean Roughness										Min-Min Roughness
Attribute	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	0.70
	1	0.80	0.85	0.70	0.75	0.85	1	1	1	0.85	

Table 3.16: MMR value

3.3.1. The partitioning attribute with the MMR is found

Table 3.15 and 3.16 show how the min-min roughness is being determined. The minimum value of roughness is taken from each attributes corresponding to each other attributes as shown in Table 3.15. Then, in table 3.16, the minimum value from all the minimum values obtain from Table 3.15 is determined, therefore 0.70 is the MMR and the attribute corresponding to it, a_4 is used as the partitioning attribute. In the case where there is more than one lowest MMR value that can be used, it is recommended to look at the next lowest MMR inside the attributes that are tied and so on until the tie is broken.

If the next lowest MMR is lower than the others, then the attribute of that corresponding MMR is selected as the partitioning attribute and binary splitting is conducted.

3.3.2. The splitting point attributes a_4 is determined

The splitting set should include the attribute value which has minimum roughness.

Taking a look at Table 3.8,

$X(a_1 = no)$ has overall minimum roughness with respect to a_i , ($i = 2, \dots, 10$) comparing to $X(a_1 = yes)$. Thus, splitting on $X(a_1 = no)$ versus $X(a_1 = yes)$ is chosen. In the case where there are more than two attributes, the splitting is on the attribute value which has the overall minimum roughness versus the other attributes. The partition at this stage can be represented as a tree and is shown in figure below.

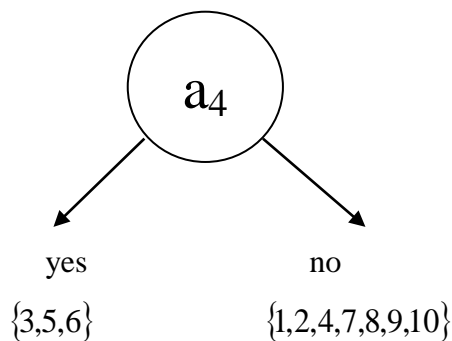


Figure 3.3: Result of clustering

3.3.3. Cluster Purity

Cluster purity is defined as the percentage of the dominant class label in each cluster.

Therefore, it can be said that, the higher the number of dominant class label in each cluster, the higher the level of the cluster's purity. Cluster purity is used as the determination of the accuracy or quality of the clusters made. However, clustering which is an unsupervised learning process does not deal with labeled or predefined dataset.

Therefore, assumptions are made where all the samples are predicted to be members of

the actual dominant class for that cluster. Mathematically, cluster purity is defined as follows:

$$Purity(i) = \frac{\text{the number of data occurring in both (i)th cluster under given threshold}}{\text{the number of data in the data set}}$$

$$Overall Purity = \frac{\sum_{i=1}^n Purity(i)}{n}$$

Where n = total number of cluster

Generally, the larger the purity value is, the better the clustering is.

CHAPTER IV

EXPECTED RESULTS AND DISCUSSION

This chapter briefly discusses on the implementation and results of rough set-based data clustering and its application to classify patients from diabetic's datasets. This chapter comprises two sections. The first section discusses the datasets used in this research, which consist of two small datasets, a real world (large) dataset and a benchmark dataset. Then the second section discusses the implementation of the rough set-based data clustering technique, namely min-min roughness (MMR) on the datasets in the first section.

4.1. Implementation

The proposed data clustering technique, namely min-min roughness (MMR) will be implemented using a software created using visual basic programming language. The dataset that is to be used need to be saved into a Microsoft Excel format file for this program to be able to read the data in the datasets. Firstly, the software will ask the user to choose a file (the Microsoft Excel file) that contains the dataset. After that is where the min-min roughness technique is implemented. The processes include in this software are: listing the indiscernibility relations of the attributes in the dataset; calculate the lower and upper approximations of each of the attribute value; calculating the roughness and

mean roughness of each attributes with respect to other attributes; determining the min-min roughness based on the mean roughness calculated; and finding the splitting point attributes which enables the user to cluster the dataset. Lastly, based on the splitting attribute found, the data will be clustered.

4.1.1. Interfaces design

There are two simple interfaces for this MMR software. One interface to browse the dataset file, and another one for all the calculations. The interfaces are as describe below:

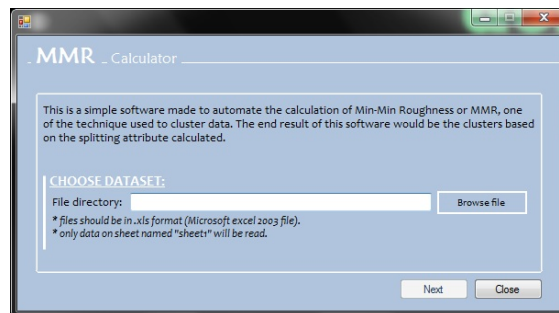


Figure 4.1: Start Interface

Figure 4.1 shows the starting interface of the MMR software. From this form, user can browse for the dataset file (Microsoft Excel file; .xls). After choosing the dataset file, user may click “Next” to continue.

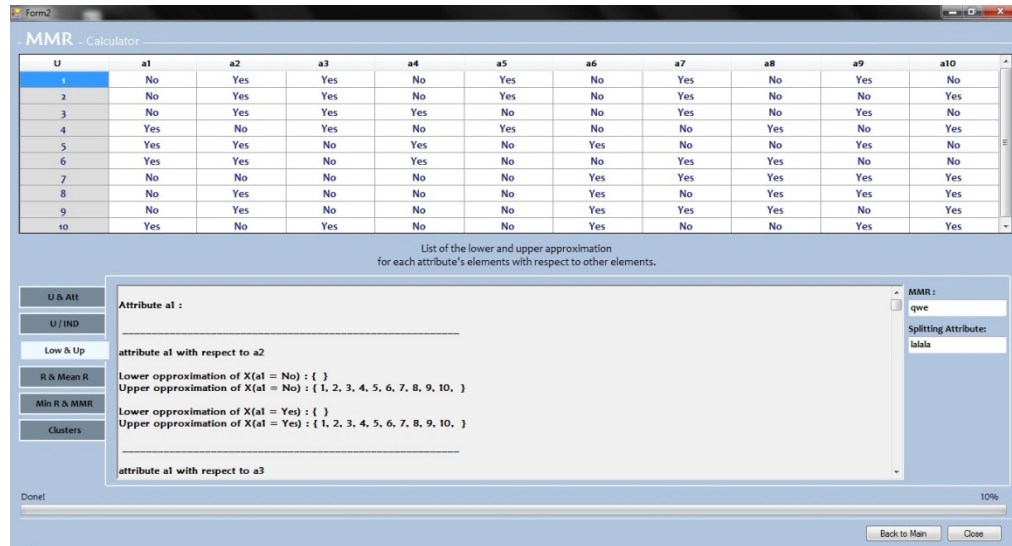


Figure 4.2: Calculation Interface

Figure 4.2 shows the interface for the calculation of the dataset that had been chosen. The top area of the interface will display the dataset that had been chosen, while the area below consists of the buttons to execute the needed calculations. The bottom middle area is the display area for the calculation result while the small textboxes at the right side is for the final result; the MMR value and the splitting attribute. A progress bar at the bottom of the interface will show how much progress of the calculation had been made.

Code for listing the Undiscernibility Relation:

```

For a = 1 To dgvData.Columns.Count - 1
    rtb2.Text = rtb2.Text & "Attribute " & dgvData.Columns(a).HeaderText &
vbNewLine & vbNewLine & vbNewLine
    UInd = ""
        For c1 = 0 To dgvData.Rows.Count - 1
            If dgvData.Item(a, c1).Tag = "True" Then
                rtb2.Text = rtb2.Text & "X(" & dgvData.Columns(a).HeaderText
& " = " & dgvData.Item(a, c1).Value & ") = { "
                UInd = UInd & "{ "
                For c2 = c1 To dgvData.Rows.Count - 1
                    If dgvData.Item(a, c2).Value = dgvData.Item(a, c1).Value
                        Then
                            UInd = UInd & dgvData.Item(0, c2).Value & ", "
                            rtb2.Text = rtb2.Text & dgvData.Item(0, c2).Value
& ", "
                        End If
                    Next
                UInd = UInd & " }, " & vbNewLine
                rtb2.Text = rtb2.Text & "}" & vbNewLine & vbNewLine
            End If
        Next
    End If

```

```

    Next
    rtb2.Text = rtb2.Text & "U/IND(" & dgvData.Columns(a).HeaderText & ")" & " =
    { " & vbNewLine & UInd & " }" & vbNewLine & vbNewLine
Next

```

Codes for Lower and Upper Approximation:

```

For a = 1 To dgvData.Columns.Count - 1
    tag = 1
    For b1 = 0 To dgvData.Rows.Count - 1
        If dgvData.Item(a, b1).Tag = "True" Then
            dgvData.Item(a, b1).Tag = CStr(tag)
            For b2 = (b1 + 1) To dgvData.Rows.Count - 1
                If dgvData.Item(a, b2).Value = dgvData.Item(a,
b1).Value Then
                    dgvData.Item(a, b2).Tag = CStr(tag)
                End If
            Next
            tag = tag + 1
        End If
    Next
Next

For a = 1 To dgvData.Columns.Count - 1

    rtb3.Text = rtb3.Text & vbNewLine & "Attribute " &
dgvData.Columns(a).HeaderText & " :" & vbNewLine & vbNewLine
    For a2 = 1 To dgvData.Columns.Count - 1
        If a <> a2 Then

            rtb3.Text = rtb3.Text & vbNewLine & "attribute " &
dgvData.Columns(a).HeaderText & " with respect to " &
dgvData.Columns(a2).HeaderText & vbNewLine

            For tag = 1 To maxAttVar(a - 1)
                For b1 = 0 To dgvData.Rows.Count - 1
                    dgvData.Item(0, b1).Tag = "None"
                Next
                For tag2 = 1 To maxAttVar(a2 - 1)
                    count2 = 0
                    For b1 = 0 To dgvData.Rows.Count - 1
                        If dgvData.Item(a2, b1).Tag = CStr(tag2) Then
                            count2 = count2 + 1
                        End If
                    Next
                    lowTestCount = 0
                    For e1 = 0 To dgvData.Rows.Count - 1
                        If dgvData.Item(a, e1).Tag = CStr(tag) Then
                            For e2 = 0 To dgvData.Rows.Count - 1
                                If dgvData.Item(a2, e2).Tag =
CStr(tag2) And dgvData.Item(0, e2).Value = dgvData.Item(0, e1).Value Then
                                    lowTestCount = lowTestCount + 1
                                End If
                            Next
                        End If
                    Next
                End If
            Next
            If lowTestCount = count2 Then

```

```

        For e2 = 0 To dgvData.Rows.Count - 1
            If dgvData.Item(a2, e2).Tag = CStr(tag2)
                Then
                    dgvData.Item(0, e2).Tag = "low"
                End If
            Next
        End If
        For e1 = 0 To dgvData.Rows.Count - 1
            If dgvData.Item(a, e1).Tag = CStr(tag) Then
                For e2 = 0 To dgvData.Rows.Count - 1
                    If dgvData.Item(a2, e2).Tag =
CStr(tag2) And dgvData.Item(0, e1).Value = dgvData.Item(0, e2).Value Then
                        For b = 0 To dgvData.Rows.Count -
1
                            If dgvData.Item(a2, b).Tag =
CStr(tag2) And dgvData.Item(0, b).Tag = "None" Then
                                dgvData.Item(0, b).Tag =
"up"
                            ElseIf dgvData.Item(a2,
b).Tag = CStr(tag2) And dgvData.Item(0, b).Tag = "low" Then
                                dgvData.Item(0, b).Tag =
"low&up"
                            End If
                        Next
                    End If
                Next
            End If
        Next
    Next
    rtb3.Text = rtb3.Text & vbNewLine & "Lower
approximation of X(" & dgvData.Columns(a).HeaderText & " = " & attVariable(a - 1,
tag - 1) & ") : { "
    attlowUpCount(a - 1, a2 - 1, 0, tag - 1) = 0
    For e1 = 0 To dgvData.Rows.Count - 1
        If dgvData.Item(0, e1).Tag = "low&up" Then
            rtb3.Text = rtb3.Text & dgvData.Item(0,
e1).Value & ", "
            attlowUpCount(a - 1, a2 - 1, 0, tag - 1) =
attlowUpCount(a - 1, a2 - 1, 0, tag - 1) + 1
        End If
    Next
    rtb3.Text = rtb3.Text & " }" & vbNewLine
    rtb3.Text = rtb3.Text & "Upper approximation of X(" &
dgvData.Columns(a).HeaderText & " = " & attVariable(a - 1, tag - 1) & ") : { "
    attlowUpCount(a - 1, a2 - 1, 1, tag - 1) = 0
    For e1 = 0 To dgvData.Rows.Count - 1
        If dgvData.Item(0, e1).Tag = "up" Or
dgvData.Item(0, e1).Tag = "low&up" Then
            rtb3.Text = rtb3.Text & dgvData.Item(0,
e1).Value & ", "
            attlowUpCount(a - 1, a2 - 1, 1, tag - 1) =
attlowUpCount(a - 1, a2 - 1, 1, tag - 1) + 1
        End If
    Next
    rtb3.Text = rtb3.Text & " }" & vbNewLine
Next
End If

```


Next

Next

Codes for Roughness and Mean Roughness Calculations:

```

For a = 1 To dgvData.Columns.Count - 1
    rtb4.Text = rtb4.Text & vbNewLine & vbNewLine &
dgvData.Columns(a).HeaderText & vbNewLine & vbNewLine
    For a2 = 1 To dgvData.Columns.Count - 1
        totalRoughness(a - 1, a2 - 1) = 0
        If a <> a2 Then
            rtb4.Text = rtb4.Text & vbNewLine & "Attribute " &
dgvData.Columns(a).HeaderText & " with respect to " &
dgvData.Columns(a2).HeaderText & vbNewLine
            For b = 1 To maxAttVar(a - 1)
                roughness(a - 1, a2 - 1, b - 1) = 1 -
(attlowUpCount(a - 1, a2 - 1, 0, b - 1) / attlowUpCount(a - 1, a2 - 1, 1, b - 1))
                rtb4.Text = rtb4.Text & "R wrt " &
dgvData.Columns(a2).HeaderText & "( X|" & dgvData.Columns(a).HeaderText & " : " &
attVariable(a - 1, b - 1) & " ) = 1 - " & CStr(attlowUpCount(a - 1, a2 - 1, 0, b
- 1)) & "/" & CStr(attlowUpCount(a - 1, a2 - 1, 1, b - 1)) & " = " &
Format(roughness(a - 1, a2 - 1, b - 1), "0.00") & vbNewLine
                totalRoughness(a - 1, a2 - 1) = totalRoughness(a - 1,
a2 - 1) + roughness(a - 1, a2 - 1, b - 1)
            Next
            meanRoughness(a - 1, a2 - 1) = totalRoughness(a - 1, a2 -
1) / maxAttVar(a - 1)
            rtb4.Text = rtb4.Text & vbNewLine & "Mean Roughness = "
& CStr(totalRoughness(a - 1, a2 - 1)) & " / " & CStr(maxAttVar(a - 1)) & " = " &
Format(meanRoughness(a - 1, a2 - 1), "0.00")
        ElseIf a = a2 Then
            meanRoughness(a - 1, a2 - 1) = 2
        End If
        progressBarIncrement()
    Next
Next

```

Next

Codes for Determining MMR value and Splitting Attribute:

```

For a = 1 To dgvData.Columns.Count - 1
    rtb5.Text = rtb5.Text & "Attribute " &
dgvData.Columns(a).HeaderText & " : Mean Roughnesses : " & vbNewLine
    first = True
    For a2 = 1 To dgvData.Columns.Count - 1
        If a <> a2 Then
            If first = True Then
                rtb5.Text = rtb5.Text & ""
                first = False
            Else
                rtb5.Text = rtb5.Text & " ; "
            End If
            rtb5.Text = rtb5.Text & Format(meanRoughness(a - 1, a2 -
1), "0.00")
        End If
        progressBarIncrement()
    Next
    rtb5.Text = rtb5.Text & vbNewLine & vbNewLine

```

```

Next
For a = 1 To dgvData.Columns.Count - 1
    For a2 = 1 To dgvData.Columns.Count - 1
        smallest = a2
        For b = a2 + 1 To dgvData.Columns.Count - 1
            If meanRoughness(a - 1, b - 1) < meanRoughness(a - 1,
smallest - 1) Then
                smallest = b
            End If
        Next
        temp = meanRoughness(a - 1, a2 - 1)
        meanRoughness(a - 1, a2 - 1) = meanRoughness(a - 1, smallest
- 1)
        meanRoughness(a - 1, smallest - 1) = temp
    Next
Next

rtb5.Text = rtb5.Text & "+++++" &
vbNewLine & vbNewLine & "Minimum mean roughnesses = " & vbNewLine

first = True
For a = 1 To dgvData.Columns.Count - 1
    If first = True Then
        rtb5.Text = rtb5.Text & ""
        first = False
    Else
        rtb5.Text = rtb5.Text & " ; "
    End If
    rtb5.Text = rtb5.Text & Format(meanRoughness(a - 1, 0), "0.00")
Next

mmr = 1
For a = 1 To dgvData.Columns.Count - 1
    If meanRoughness(a - 1, 0) < mmr Then
        mmr = meanRoughness(a - 1, 0)
        minColumn(0) = a
    End If
Next
minCount(0) = 0
For a = 1 To dgvData.Columns.Count - 1
    If mmr = meanRoughness(a - 1, 0) Then
        minCount(0) = minCount(0) + 1
        ReDim Preserve minColumn(minCount(0))
        minColumn(minCount(0) - 1) = a
    End If
Next
For a2 = 2 To dgvData.Columns.Count - 1
    If minCount(a2 - 2) > 1 Then
        mmr = 1
        For b = 1 To minCount(a2 - 2)
            If meanRoughness(minColumn(b - 1) - 1, a2 - 1) < mmr Then
                rtb6.Text = rtb6.Text & minColumn(b - 1) & " | "
                mmr = meanRoughness(minColumn(b - 1) - 1, a2 - 1)
            End If
        Next
        minCount(a2 - 1) = 0
        For b = 1 To minCount(a2 - 2)
            If mmr = meanRoughness(minColumn(b - 1) - 1, a2 - 1) Then

```

```

        minCount(a2 - 1) = minCount(a2 - 1) + 1
        minColumn(minCount(a2 - 1) - 1) = minColumn(b - 1)
    End If
Next
ReDim Preserve minColumn(minCount(a2 - 1))
End If
Next

txtMMR.Text = Format(mmr, "0.00")
txtSA.Text = dgvData.Columns(minColumn(0)).HeaderText

```

Codes for Clustering Based on the Splitting Attribute Found:

```

For a = 0 To dgvData.Rows.Count - 1
    dgvData.Item(minCol, a).Tag = "True"
Next
For a = 0 To dgvData.Rows.Count - 1
    If dgvData.Item(minCol, a).Tag = "True" Then
        rtb6.Text = rtb6.Text & "-----> { " & dgvData.Item(0, a).Value
        & ", "
        For b = a + 1 To dgvData.Rows.Count - 1
            If dgvData.Item(minCol, a).Value = dgvData.Item(minCol,
            b).Value Then
                rtb6.Text = rtb6.Text & dgvData.Item(0, b).Value & ", "
                dgvData.Item(minCol, b).Tag = "False"
            End If
        Next
        rtb6.Text = rtb6.Text & "}" & vbNewLine
    End If
Next

```

4.2. Datasets

The MMR software mentioned above will be used test the MMR technique that on three different datasets. The three datasets are describe below:

4.2.1. Small datasets

The first dataset is a dataset made by picking random data from the dataset of diabetics of obtained Hospital Ampuan Afzan. From the large dataset, 10 attributes, which are the symptom of diabetes, are chosen and 10 patients with their respective attributes are chosen randomly. The attributes of this small dataset are: often thirst; excessive hunger; frequent urination; tiredness and fatigue; rapid and/or sudden weight loss; blurred vision;

numbness and/or tingling in the hands and feet; slow healing of minor-to-treat yeast infection in women; recurrent or hard-to-treat yeast infection in women; and dry or itchy skin.

The second small dataset is the dataset created randomly from the large dataset of Pima Indians Diabetes Database [1]. From the large dataset, 10 diabetes patients are picked randomly with all their respective 8 attributes to create the small dataset. Those attributes are: number of times pregnant; plasma glucose concentration a 2 hours in an oral glucose tolerance test; diastolic blood pressure; triceps skin fold thickness; 2-hour serum insulin; body mass index; diabetes pedigree function; and age.

4.2.2. Large dataset (real world dataset)

The main dataset for this research is a real world dataset obtain from the patients of Hospital Ampuan Afzan, Kuantan. There are a total of 252 diabetics in this dataset. The attributes of this dataset are the symptoms of diabetes, same as the ones mentioned in 4.1.1 but added with another one attributes which is the decision (type of diabetes, either type 1 or type 2).

4.2.3. Benchmark dataset

The benchmark dataset for this research is the dataset of Pima Indians Diabetes Database [1]. Particularly, all patients in this dataset are females at least 21 years old of Pima Indian heritage. There are 768 patients in this dataset and 8 attributes, the same as the ones mentioned in 4.1.1 (9 including class variables).

CHAPTER V

CONCLUSION

Min-min roughness technique is a useful technique to cluster large dataset with multi-valued attribute and with no clear structure or classes, which exist in the real world. It relates the data by analyzing the indiscernibility relation between the data to enable the data to be clustered properly. To have a proper or meaningful cluster is useful for fields such as machine learning, data mining, pattern recognition, image analysis, and bioinformatics. In this project, the proposed min-min roughness is applied to the dataset of diabetic patients. This technique is based on the rough set theory by Pawlak in 1981. The uses of this theory and implementation of it had been presented in the previous chapters. Firstly the indiscernibility for every data in the dataset is calculated and the roughness of each data is determined. Based on this roughness, the minimum roughness can be obtained which determined the clustering point for the dataset. By implementing this theory into the diabetic dataset, the dataset can be split into meaningful clusters, which will be useful for further research on the dataset or on diabetes symptoms.

REFERENCES

- 10 Facts About Diabetes, <http://www.who.int/features/factfiles/diabetes/facts/en/index.html>, Retrieved October 10, 2011.
- Ahmad A., Dey L., A k-mean clustering algorithm for mixed numeric and categorical data, *Data & Knowledge Engineering*, Vol. 63, Issue 2, pp. 503-527.
- All About Diabetes, <http://www.medicalnewstoday.com/info/diabetes/>, Retrieved October 9, 2011.
- Asian Diabetes Drug Tests, <http://www.asiandiabetes.org/category/asian-diabetes-statistics/>, Retrieved October 10, 2011.
- Boffetta P., McLerran D., Chen Y., Inoue M., Sinha R., He J., Gupta P.C., Tsugane S., Irie F., Tamakoshi A., Gao Y.T., Shu X.O., Wang R., Tsuji I., Kuriyama S., Matsuo K., Satoh H., Chen C.J., Yuan J.M., Yoo K.Y., Ahsan H., Pan W.H., Gu D., Pednekar M.S., Sasazuki S., Sairenchi T., Yang G., Xiang Y.B., Nagai M., Tanaka H., Nishino Y., You S.L., Koh W.P., Park S.K., Shen C.Y., Thornquist M., Kang D., Rolland B., Feng Z., Zheng W., Potter J.D., Body Mass Index and Diabetes in Asia: A Cross-Sectional Pooled Analysis of 900,000 Individuals in the Asia Cohort Consortium, *PloS one*, Vol.6, No. 6.
- Cattral R., Oppacher F., Deugo, D., Supervised and unsupervised data mining with an evolutionary algorithm, *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, Vol.2, No., pp.767-774.
- Chang I. C., Hsiao S. J., Hsu H. M., Chen T. H., Building mPHR to assist diabetics in self-healthcare management, *Service Systems and Service Management (ICSSSM), 2010 7th International Conference on*, pp.1-5, 28-30 June 2010
- Chen D., Cui D. W., Wang C. X., Wang Z. R., A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data, *International Journal of Information Technology*, vol. 12, No.3, 200
- Chen M.S., Han J., Yu P.S., Data Mining: An Overview from Database Perspective, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, Issue 6, pp. 866-883

Cheong C. W., Clustering and Predicting Stock Prices Using Advanced Wavelet Methods and Neural Networks – a study on KLSE

Chimphlee S., Salim N., Ngadiman M. S., Chimphlee W., Srinoy S., Rough Sets Clustering and Markov model for Web Access Prediction, *In Postgraduate Annual Research Seminar 2006 (PARS 2006)*, 24 - 25 Mei 2006.

Clustering, http://www.coli.uni-saarland.de/~crocker/Teaching/Connectionist/lecture13_4up.pdf, Retrieved October 19, 2011

Colin L. Carter and Howard J. Hamilton. Performance Evaluation of Attribute-Oriented Algorithms for Knowledge Discovery from Databases. In the Proceeding of Seventh International Conference on [Tools with Artificial Intelligence](#), 1995, 486-489.

Computer Science, http://en.wikipedia.org/wiki/Computer_science, Retrieved October 17, 2011

Diabetes among children on the rise, <http://thestar.com.my/news/story.asp?file=/2010/7/24/nation/6710240&sec=nation>, retrieved October 14, 2011

Diabetes hits top ranks in Malaysia, <http://www.asiaone.com/Health/News/Story/A1Story20100802-229924.html>, Retrieved October 15, 2011

Diabetes Mellitus, http://en.wikipedia.org/wiki/Diabetes_mellitus, Retrieved October 14, 2011

Diabetes, http://www.emedicinehealth.com/diabetes/article_em.htm#Diabetes, Overview, Retrieved October 9, 2011.

Diabetes, <http://www.who.int/mediacentre/factsheets/fs312/en/>, Retrieved October 9, 2011.

Fayyad U., Piatetsky-Shapiro G., Smyth P., From Data Mining to Knowledge Discovery in Databases *Ai Magazine*, Vol. 17, No. 3, pp. 37-54

Fung G., A Comprehensive Overview of Basic Clustering Algorithms, *IEEE June*, pp. 1-37

- Halkidi M., Batistakis Y., Vazirgiannis M., On Clustering Validation Technique, *J. Intell. Inf. Sys.*, Vol. 17, Issue 2-3, pp. 107-145
- Han J., Chiang J. Y., Chee S., Chen J., Chen Q., Cheng S., Gong W., Kamber M., Koperski K., Liu G., Lu Y., Stefanovic N., Winstone L., Xia B. B., Zaiane O. R., Zhang S., Zhu H., DBMiner: A System for Data Mining in Relational Databases and Data Warehouses, in *CASCON '97 Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative research*, 1997.
- He Z., Xu X., Deng S., A cluster ensemble method for clustering categorical data, *Information Fusion*, Vol. 6, No. 2, pp. 143-151.
- Jain A. K., Murty M. N., Flynn P.J., Data Clustering: A Review, *ACM Computing Surveys (CSUR)*, Vol. 31, Issue 3, pp. 264-323.
- Kumar P., Krishna P. R., Bapi R. S., De S. K., Rough clustering of sequential data, *Data & Knowledge Engineering*, Vol. 63, Issue 2, pp. 183-199.
- Lazăr A., Research Professorship Proposal - Knowledge Discovery for Large Data Sets
- Letchuman G.R., Wan Nazaimoon W.M., Wan Mohamad W.B., Chandran L.R., Tee G.H., Jamaiyah H., Isa M.R., Zanariah H., Fatanah I., Ahmad Faudzi Y., Prevalence of Diabetes in the Malaysian National Health Morbidity Survey III 2006, *Med J Malaysia*, pp. 180-6.
- Linden R., Clustering Numerical and Categorical Data, Faculdade Salesiana Maria Auxiliadora, RJ, Brazil, 2004, pp. 1-3
- M. Magnani, Technical report on Rough Set Theory for Knowledge Discovery in Data Bases. Technical Report. University of Bologna, Department of Computer Science, 2003.
- Math Lesson 7: Two Types of Data – Numerical (Quantitative) and Categorical (Qualitative), <http://www.malamaaina.org/files/mathematics/lesson7.pdf>, Retrieved October 20, 2011.
- Mazlack L., He A., Zhu Y., Coppock S., A rough set approach in choosing partitioning attributes, in *Proceedings of the ISCA 13th International Conference (CAINE-2000)*, 2000, pp. 1-6.
- Mitra, S., Pal, S.K., Mitra, P., Data Mining in Soft Computing Framework: A Survey, *IEEE Transactions on Neural Networks*, Vol. 13, Issue 1, pp. 3-14.

- Nguyen H. S., Skowron A., Rough Set Approach to KDD, *In Proceedings of the 3rd international conference on Rough sets and knowledge technology*. pp. 19-20
- Parmar D., Wu T., Callerman T., Fowler J. W., Wolfe P., A clustering algorithm for supplier base management, *International Journal of Production Research*, Vol. 48, No. 13, January 2010, pp.3802-3821.
- Peters M., Zaki M.J., Click: Clustering categorical data using k-partite maximal cliques, *International Conference on Data Engineering (ICDE)*, 2005.
- Prediabetes FAQs, <http://www.diabetes.org/diabetes-basics/prevention/pre-diabetes/pre-diabetes-faqs.html>, Retrieved October 23, 2011.
- Ralambondrainy H., A conceptual version of the K-means algorithm, *Pattern Recognition Letters*, Volume 16, Issue 11, pp. 1147-1157.
- Ramachandran A., Ma R.C., Snehalatha C., Diabetes in Asia, *The Lancet*, Volume 375, pp 408-418.
- Rashid T., Clustering of Fuzzy Image Features, MSc Thesis, 2003
- Spatial Data, http://www.webopedia.com/TERM/S/spatial_data.html, Retrieved October 20, 2011.
- Symptoms of Diabetes, <http://www.informationaboutdiabetes.com/symptoms-of-diabetes>, Retrieved October 10, 2011.
- Top 7 Risk Factors for Type 2 Diabetes, <http://diabetes.about.com/od/symptomsdiagnosis/tp/riskfactors.htm>, Retrieved October 23, 2011.
- Tripathy B. K., Ghosh A., SSSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory, *Advances in Applied Science Research*, 2011, Vol. 2, No. 3, pp. 314-326.
- Tripathy B.K., Tripathy H.K. and Das P.K., An Intelligent Approach of Rough Set in Knowledge Discovery Databases, *Proceedings of World Academy of Science, International Journal of Computer Science and Engineering* Vol. 2, No. 1, pp. 45-48.
- Types of Diabetes, http://www.healthinsite.gov.au/topics/Types_of_Diabetes, Retrieved October 9, 2011.

- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In the Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD) 1996, 83-88.
- Yao Y. Y., A Comparative Study of Fuzzy Sets and Rough Sets, *Information Sciences: an International Journal*, Vol. 109, Issue 1-4, August 1998, pp. 227-242.
- Yen S. J., Lee Y. S., An incremental data mining algorithm for discovering web access pattern, *International Journal of Business Intelligence and Data Mining* 2006, Vol. 1, No.3 pp. 288-303.
- Zadeh L. A., Fuzzy Sets, *Information and Control*, Vol 8, Issue 3, pp. 338-353.
- Zaini A., Where is Malaysia in the midst of the Asian epidemic of diabetes mellitus?, *Diabetes Research and Clinical Practice*, Vol. 50, Supplement 2, pp. S23-S28.