

ROUGH SET APPROACH FOR CATEGORICAL DATA CLUSTERING

TUTUT HERAWAN

A thesis submitted in fulfillment of  
the requirements for the award of the  
Doctor of Philosophy



Faculty of Information Technology and Multimedia  
Universiti Tun Hussein Onn Malaysia

MARCH 2010

## ABSTRACT

A few techniques of rough categorical data clustering exist to group objects having similar characteristics. However, the performance of the techniques is an issue due to *low accuracy, high computational complexity* and *clusters purity*.

This work proposes a new technique called *Maximum Dependency Attributes* (MDA) to improve the previous techniques due to these issues. The proposed technique is based on rough set theory by taking into account the dependency of attributes of an information system. The main contribution of this technique is to introduce a new technique to classify objects from categorical datasets which has better performance as compared to the baseline techniques.

The algorithm of the proposed technique is implemented in MATLAB® version 7.6.0.324 (R2008a). They are executed sequentially on a processor Intel Core 2 Duo CPUs. The total main memory is 1 Gigabyte and the operating system is Windows XP Professional SP3. Results collected during the experiments on four small datasets and thirteen UCI benchmark datasets for selecting a clustering attribute show that the proposed MDA technique is an efficient approach in terms of accuracy and computational complexity as compared to BC, TR and MMR techniques. For the clusters purity, the results on Soybean and Zoo datasets show that MDA technique provided better purity up to 17% and 9%, respectively.

The experimental result on supplier chain management clustering also demonstrates how MDA technique can contribute to practical system and establish the better performance for computation complexity and clusters purity up to 90% and 23%, respectively.

## ABSTRAK

Terdapat beberapa teknik perkelompokan data berkategori kasar yang digunakan untuk mengumpulkan objek yang mempunyai ciri-ciri yang sama. Walaubagaimana pun, prestasi teknik-teknik ini mempunyai isu daripada segi ketepatan yang rendah, kekompleksan yang tinggi dan kelompok yang mempunyai ketulenan rendah.

Satu teknik baru, iaitu *Maximum Dependency Attributes* (MDA) dicadangkan untuk memperbaiki teknik-teknik lalu berkaitan dengan isu-isu tersebut. Teknik yang dicadangkan adalah berdasarkan kepada set teori kasar dengan mengambil kira kebergantungan atribut dalam sistem maklumat. Sumbangan utama teknik itu ialah ia dapat mengklasifikasikan objek dari set data berkategori yang mempunyai prestasi yang lebih baik berbanding dengan teknik '*baseline*'.

Algoritma teknik yang dicadangkan telah diimplementasi menggunakan MATLAB® versi 7.6.0.324 (R2008a). Ia telah dilarikan menggunakan pemproses Intel Core 2 Duo yang mempunyai 1GB memori dan sistem pengoperasian Windows XP Professional SP3. Keputusan eksperimen ke atas empat set data kecil dan tigabelas set data penanda aras UCI untuk pemilihan atribut kelompok menunjukkan bahawa teknik MDA yang dicadangkan adalah merupakan satu teknik yang efisien dari sudut ketepatan dan kekompleksan, jika dibandingkan kepada BC, TR dan MMR. Untuk ketulenan kelompok, keputusan eksperimen ke atas set data Soybean dan Zoo menunjukkan teknik MDA memberikan ketulenan yang lebih baik iaitu sehingga 17% dan 19%.

Keputusan eksperimen ke atas kelompok rangkaian pengurusan pembekal juga menunjukkan bagaimana teknik MDA boleh memberi sumbangan kepada sistem praktikal dan memberikan prestasi yang lebih baik kepada kekompleksan dan ketulenan kelompok sehingga 90% dan 23%.

## PUBLICATIONS

A fair amount of the materials presented in this thesis has been published in various refereed conference proceedings and journals.

1. Tutut Herawan and Mustafa Mat Deris, Rough set theory for topological space in information systems, *The Proceeding of International Conference of AMS'09*, IEEE Press, Pages 107–112, 2009.
2. Tutut Herawan and Mustafa Mat Deris, A construction of nested rough set approximation in information systems using dependency of attributes, *The Proceeding of International Conference of PCO 2009*, American Institute of Physic 1159, Pages 324–331, 2009.
3. Tutut Herawan and Mustafa Mat Deris, Rough topological properties of set in information systems using dependency of attributes, *The Proceeding of International Conference of CITA '09*, Pages 39–46, 2009.
4. Tutut Herawan and Mustafa Mat Deris, A framework on rough set-based partitioning attribute selection, *Lecture Notes in Artificial Intelligence Volume 5755*, Part 1, Springer Verlag, Pages 91–100, 2009.
5. Tutut Herawan, Iwan Tri Riyadi Yanto and Mustafa Mat Deris, Rough set approach for categorical data clustering, *Communication of Computer and Information Sciences Volume 64*, Springer Verlag, Pages 188–195, 2009.

6. Tutut Herawan, Iwan Tri Riyadi Yanto and Mustafa Mat Deris, A construction of hierarchical rough set approximations in information systems using dependency of attributes, *Studies in Computational Intelligence* Volume 283, Springer Verlag, Pages 3–15, 2010.
7. Tutut Herawan, Mustafa Mat Deris and Jemal H. Abawajy, Rough set approach for selecting clustering attribute, *Knowledge Based Systems*, Elsevier, Volume 23, Issue 3, Pages 220–231, April 2010.
8. Tutut Herawan, Rozaida Ghazali, Iwan Tri Riyadi Yanto and Mustafa Mat Deris, Rough clustering for categorical data, *International Journal of Database Theory and Application*, a special issue of DTA 2009, Volume 3, Issue 1, March 2010, 33–52.
9. Tutut Herawan, Iwan Tri Riyadi Yanto and Mustafa Mat Deris, ROSMAN: ROugh Set approach for clustering supplier chain MANagement, Manuscript accepted on special issue of Soft Computing Methodology, *International Journal of Biomedical and Human Sciences*, Japan, to appear in Volume 17, Issue 1, July 2010.



**CONTENTS**

<b>TITLE</b>	<b>i</b>
<b>DECLARATION</b>	<b>ii</b>
<b>ACKNOWLEDGEMENT</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>iv</b>
<b>ABSTRAK</b>	<b>v</b>
<b>PUBLICATIONS</b>	<b>vi</b>
<b>CONTENTS</b>	<b>viii</b>
<b>LIST OF TABLES</b>	<b>xi</b>
<b>LIST OF FIGURES</b>	<b>xiii</b>
<b>LIST OF APPENDICES</b>	<b>xiv</b>

**CHAPTER 1 INTRODUCTION**

1.1	Background	1
1.3	Problems Statement	4
1.4	Objectives and Scope	6
1.5	Contributions	7
1.6	Thesis Organization	7

**CHAPTER 2 ROUGH SET THEORY**

2.1	Information System	10
2.2	Indiscernibility Relation	12
2.3	Approximation Space	13
2.4	Set Approximations	14
2.5	Summary	15

**CHAPTER 3 CATEGORICAL DATA CLUSTERING USING ROUGH SET THEORY**

3.1	Data Clustering	20
3.2	Categorical Data Clustering	22
3.3	Categorical Data Clustering using Rough Set Theory	24
3.3.1	The TR Technique	25
3.3.2	The MMR Technique	27
3.3.3	The Relation between TR and MMR Techniques	28
3.3.4	Summary	37

**CHAPTER 4 MAXIMUM DEPENDENCY OF ATTRIBUTES (MDA) TECHNIQUE**

4.1	Dependency of Attributes in an Information System	38
4.2	Selecting a Clustering Attribute	40

4.2	The Accuracy	43
4.3	The Computational Complexity	46
4.4	Objects Partitioning	46
4.5	Clusters Validation	47
4.6	Summary	48

## **CHAPTER 5 EXPERIMENTAL RESULTS**

5.1	Comparisons for Selecting a Clustering Attribute	49
5.1.1	The Credit Card Promotion Dataset	50
5.1.2	The Student's Enrollment Qualifications Dataset	52
5.1.3	Animal World Dataset	58
5.1.4	A Dataset from Parmar <i>et. al.</i>	63
5.1.5	The Benchmark and Real World Datasets	68
5.2	Clusters Validation	72
5.2.1	Soybean Dataset	72
5.2.2	Zoo Dataset	73
5.3	Application in Clustering Supplier Chain Management	76
5.4	Summary	81

## **CHAPTER 6 CONCLUSION AND FUTURE WORKS**

6.1	Conclusion	82
6.2	Future Works	83

<b>REFERENCES</b>	85
-------------------	----

<b>APPENDIX</b>	96
-----------------	----

<b>VITAE</b>	100
--------------	-----



## LIST OF TABLES

2.1	An information system	10
2.2	A student decision system	11
3.1	The credit card promotion dataset	30
3.2	The total roughness of attributes from Table 3.1	33
3.3	The minimum-minimum roughness of attributes from Table 3.1	35
3.4	The modified MMR of all attributes from Table 3.1	36
4.1	A modified information system from (Pawlak, 1983)	44
4.2	The degree of dependency of attributes from Table 4.1	45
5.1	The degree of dependency of attributes from Table 3.1	50
5.2	The student's enrollment qualifications dataset	53
5.3	The total roughness of attributes from Table 5.2	54
5.4	The minimum-minimum roughness of attributes from Table 5.2	55
5.5	The degree of dependency of attributes from Table 5.2	56
5.6	Animal world dataset	58
5.7	The total roughness of attributes from Table 5.6	59
5.8	The minimum-minimum roughness of attributes from Table 5.6	60
5.9	The degree of dependency of attributes from Table 5.6	61
5.10	A dataset from Parmar <i>et. al.</i>	63
5.11	The total roughness of attributes from Table 5.10	64
5.12	The minimum-minimum roughness of attributes from Table 5.10	65
5.13	The degree of dependency of attributes from Table 5.10	66
5.14	The benchmark and real world datasets	69
5.15	The clusters purity of soybean dataset from MDA	73
5.16	The clusters purity of soybean dataset from MMR	73

5.17	The clusters purity of zoo dataset from MDA	74
5.18	The clusters purity of zoo dataset from MMR	75
5.19	The overall improvement of clusters purity from MMR by MDA	76
5.20	A discretized supplier dataset	77
5.21	The clusters purity of supplier dataset from MDA	80
5.22	The clusters purity of supplier dataset from MMR	80
5.23	The overall improvement of clusters purity from MMR by MDA	81



## LIST OF FIGURES

2.1	Set approximations	15
4.1	The MDA algorithm	43
4.2	The objects splitting	47
5.1	The accuracy of BC, TR, MMR and MDA from case 5.1.1	51
5.2	The computational complexity of BC, TR, MMR and MDA from case 5.1.1	52
5.3	The accuracy of BC, TR, MMR and MDA from case 5.1.2	57
5.4	The computational complexity of BC, TR, MMR and MDA from case 5.1.2	57
5.5	The accuracy of BC, TR, MMR and MDA from case 5.1.3	62
5.6	The computational complexity of BC, TR, MMR and MDA from case 5.1.3	62
5.7	The accuracy of TR, MMR and MDA from case 5.1.4	67
5.8	The computational complexity of TR, MMR and MDA from case 5.1.4	67
5.9	The accuracy of BC, TR, MMR and MDA from case 5.1.5	70
5.10 A	The executing time of BC, TR, MMR and MDA from case 5.1.5	70
5.10 B	The executing time of BC, TR, MMR and MDA from case 5.1.5	71
5.10 C	The executing time of BC, TR, MMR and MDA from case 5.1.5	71
5.11	The comparison of overall clusters purity	75
5.12	The computation of MDA and MMR of SCM dataset	79
5.13	The executing time of SCM dataset	79

## LIST OF APPENDICES

MATLAB codes for TR, MMR and MDA Techniques 96

VITAE 100



PTTA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH

## CHAPTER 1

### INTRODUCTION

In this section, the background of the research is outlined, followed by problem statements, the objectives and the scope of the research, contributions and lastly, the thesis organization.

#### 1.1 Background

It is estimated that every 20 months or so the amount of information in the world doubles. In the same way, tools for use in the various knowledge fields (acquisition, storage, retrieval, maintenance, and etc) must be developed to combat this growth (Jensen, 2005). Due to the explosion of data in the modern society, most organizations have large databases that contain a wealth of undiscovered, yet valuable information. Because the amount of data is so huge that it is usually very difficult to examine these data by human eyes to discover knowledge of our interest. This leads to a significant research on knowledge discovery process, particularly knowledge discovery in databases (KDD) (Piatesky-Shapiro, Fayyad and Smyth, 1996; Sever, 1998; Dutsch and Gediga, 2000; Atkinson-Abutridy, Mellish and Aitken, 2004). Knowledge discovery in databases is a process of discovering

previously unknown, valid, novel, potentially useful and understandable patterns in large datasets (Piatesky-Shapiro, Fayyad and Smyth, 1996). The KDD process can be decomposed into the following steps:

a. Data Selection:

A target dataset is selected or created. Several existing datasets may be joined together to obtain an appropriate example set.

b. Data Cleaning/Preprocessing:

This phase includes, among other tasks, noise removal/reduction, missing value imputation, and attribute discretization. The goal of this is to improve the overall quality of any information that may be discovered.

c. Data Reduction:

Most datasets will contain a certain amount of redundancy that will not aid knowledge discovery and may in fact mislead the process. The aim of this step is to find useful features to represent the data and remove non-relevant ones. Time is also saved during the data mining step as a result of this.

d. Data Mining:

A data mining method (the extraction of hidden predictive information from large databases) is selected depending on the goals of the knowledge discovery task. The choice of algorithm used may be dependent on many factors, including the source of the dataset and the values it contains.

e. Interpretation/Evaluation:

Once knowledge has been discovered, it is evaluated with respect to validity, usefulness, novelty and simplicity. This may require repeating some of the previous steps (Piatesky-Shapiro, Fayyad and Smyth, 1996).

The fourth step in the knowledge discovery process, namely data mining, is the process of extracting patterns from data. Data mining encompasses many different techniques and algorithms, including classification, clustering, association rule, and so on.

There are two distinct areas of data mining: supervised data mining and unsupervised data mining. Both of these areas exploit the techniques such as of subset formation/selection and granulizing of continuous features. Supervised data mining techniques usually determine in advance the subjects that are of interest. These techniques use a well-defined (known) dependent variable. This greatly limited the searching space and is proved to fast and efficient in the data mining

process. Regression and classification techniques are examples of supervised methods. But because the subject of interest is pre-determined, this is counter-intuitive to the general goal of conducting mining to find unexpected, interesting things (Mazlack, 1996). Another data mining approach is to use unsupervised methods that use non-semantic heuristics rather than pre-determined subject of interest. The knowledge supplied to these systems only includes the syntactic characteristics of the database. In these systems, grouping of the data are defined without the use of a dependent variable. Therefore, this approach can be applied to more than one semantic domain. Heuristics are often adopted from the following areas: information theory (Quinlan, 1986; Agrawal, Imielinski and Swami, 1993; Zadeh, 1965; Pawlak, 1982; Molodtsov, 1999) and statistics (Shen, 1991; Langley, Iba and Thompson, 1992). Unsupervised mining has difficult design concerns. The main difficulty is combinational explosion. Supervised search can use domain knowledge to reduce the search space. However, only heuristics are available for unsupervised search.

One of the unsupervised data mining techniques is based on rough set theory, a mathematical formalism developed by Z. Pawlak to analyze data tables (Pawlak, 1982; Pawlak, 1991; Pawlak and Skowron, 2007; Pawlak and Skowron, 2007). Its peculiarity is a well understood formal model, which allows to find several kinds of information, such as relevant features or classification rules. The application of rough set theory for data mining is one approach that has proved successfully (Magnani, 2005). Over the past twenty years, rough set theory has become a topic of great interest to researchers and has been applied to many domains, such as data classification (Chouchoulas and Shen, 2001), data clustering (Parmar, Wu, Callarman, Fowler and Wolfe, 2009; Yanto, Herawan and Mat Deris, 2010a; Yanto, Herawan and Mat Deris, 2010b; Yanto, Herawan and Mat Deris, 2010c) and association rules mining (Guan, Bell and Liu, 2003; Bi, Anderson and McClean, 2003; Guan, Bell and Liu, 2005). This success due to the main advantage of rough set theory in data mining, i.e., it does not needs any preliminary or additional information about data, like probability in statistics or grade of membership in fuzzy set theory (Zadeh, 1965).

## 1.2 Problems Statement

Since classification is the philosophy of classical rough set theory, i.e. rough set theory was used mainly to classify objects or to assign them to classes known as *a posteriori* (Komorowski, Polkowski and Skowron, 1999). Therefore, this thesis focuses on application of rough set theory for data clustering (*a priori*), particularly, for categorical data clustering.

Clustering a set of objects into homogeneous classes is a fundamental operation in data mining. The operation is required in a number of data analysis tasks, such as unsupervised classification and data summation, as well as segmentation of large homogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed (Halkidi, Batistakis and Vazirgiannis, 2001).

Cluster analysis techniques have been used in many areas such as manufacturing, medicine, nuclear science, radar scanning and research and development planning. For example, Haimov *et al.* use cluster analysis to segment radar signals in scanning land and marine objects (Haimov, Michalev, Savchenko and Yordanov, 1989). Wong *et al.* present an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET) (Wong, Feng, Meikle and Fulham, 2002). Jiang *et al.* analyzed a variety of cluster techniques for complex gene expression data (Jiang, Tang and Zhang, 2004). Wu *et al.* develop a clustering algorithm specifically designed to handle the complexities of gene data that can estimate the correct number of clusters and find them (Wu, Liew, Yan and Yang, 2004). Mathieu and Gibson use cluster analysis as a part of a decision support tool for large-scale research and development planning to identify programs to participate in and to determine resource allocation (Mathieu and Gibson, 2004). Saglam *et al.* proposed a mathematical programming based clustering approach that is applied to a digital platform company's customer segmentation problem involving demographic and transactional attributes related to the customers. The clustering problem is formulated as a mixed-integer programming problem with the objective of minimizing the maximum cluster diameter among all clusters (Saglam, Salman, Sayın and Türkay, 2006). Fathian *et al.* proposed a hybridization of nature inspired intelligent technique with K-means algorithm. The



HBMK-Means (honeybee mating K-means) is proposed to improve the K-Means technique (Fathian, Amiri and Maroosi, 2007). Finally, Cheng and Leu proposed an effective clustering algorithm, the constrained k-prototypes (CKP) algorithm is proposed to resolve the classification problems of construction management (Cheng and Leu, 2009).

A problem with many of the clustering methods and applications mentioned above is that they are applicable for clustering data having numerical values for attributes. Those works in clustering are focused on attributes with numerical value due to the fact that it is relatively easy to define similarities from the geometric position of the numerical data.

Currently, more attentions of clustering techniques have been put on categorical data. Unlike numerical data, categorical data have multi-valued attributes. Thus, similarity can be defined as common objects, common values for the attributes, and the association between the two. In such cases, the horizontal co-occurrences (common attributes for the objects) as well as the vertical co-occurrences (common values for the attributes) can be examined (Wu, Liew, Yan and Yang, 2004). A number of algorithms for clustering categorical data have been proposed including work by Dempster *et al.* (Dempster, Laird and Rubin, 1997), Ganti *et al.* (Ganti, Gehrke, Ramakrishnan, 1999), Gibson *et al.* (Gibson, Kleinberg and Raghavan, 2000), Guha *et al.* (Guha, Rastogi and Shim, 2000), Zaki *et al.* (Zaki, Peters, Assent and Seidl, 2007), and Chen and Liu (Chen and Liu, 2009). While these methods make important contributions to the issue of clustering categorical data, they are not designed to handle uncertainty in the clustering process. This is an important issue in many real world applications where there is often no sharp boundary between clusters. Therefore, there is a need for a robust clustering algorithm that can handle uncertainty in the process of clustering categorical data.

One of the data clustering techniques is based on rough set theory. The main idea of the rough clustering is the clustering data set is mapped as the decision table and this can be done by introducing a decision attribute. Currently, there has been work in the area of applying rough set theory in the process of selecting clustering attribute. Mazlack *et al.* proposed two techniques to select clustering attribute: i.e., Bi-Clustering (BC) technique based on bi-valued attributes and Total Roughness (TR) technique based on the average of the accuracy of approximation (accuracy of roughness) in the rough set theory (Mazlack, He, Zhu and Coppock,

2000). Parmar *et al.* proposed a new technique called Min-Min Roughness (MMR) for selecting clustering attribute to improve BC technique for data set with multi-valued attributes (Parmar, Wu and Blackhurst, 2007). However, since the algorithm for categorical data clustering based on rough set theory is relatively new, the focus of MMR algorithm has been on evaluating the performance. In reviewing BC, TR and MMR techniques, we point out their drawbacks as follows:

- a. The issue of accuracy is faced to BC technique, since it selects a clustering attribute based on bi-valued attributes without further calculation of accuracy of approximations.
- b. For TR and MMR techniques, due to all attributes are considered to be selected and the ever-increasing computing capabilities, computation complexity is still be an outstanding issue.
- c. For cluster validity, the clusters purity of MMR is still an issue due to objects in different class appeared in a cluster.

Therefore, based on these drawbacks, there is a need for improving of those techniques. In this work, a technique termed MDA (Maximum Dependency of Attributes) for categorical data clustering aimed to mine the hidden “nuggets” patterns in database is proposed. It is based on rough set theory taking into account maximal dependency of attributes in an information system. Experimental tests on small datasets, benchmark datasets and real world datasets demonstrate how such techniques can contribute to practical system, such as for supplier chain management clustering.

### 1.3 Objectives and Scope

This research embarks on the following objectives:

- a. To develop a technique for clustering categorical data using rough set theory based on dependency of attributes having the ability to achieve higher accuracy, lower computation complexity and higher clusters purity.
- b. To elaborate the goals in the proposed technique on small datasets, benchmark datasets and real world datasets.

- c. To do a comparison between the proposed technique with the baseline techniques based on accuracy, computation complexity and clusters purity.

The scope of this research falls within categorical data clustering using rough set theory.

#### 1.4 Contributions

The specific contributions of this thesis correspond to the three factors as described earlier, which are:

- a. Increasing accuracy in selecting a clustering attribute.
- b. Reducing complexity in selecting a clustering attribute.
- c. Increasing clusters purity in classifying objects.

#### 1.5 Thesis Organization

The rest of this thesis is organized as follows:

Chapter 2 describes the fundamental concept of rough set theory. The notion of an information system and its relation with a relational database, the concept of an indiscernibility relation induced by a subset of the whole set of attributes, the concept of a (Pawlak) approximation space, the notion of set approximations and its quality of approximations are described.

Chapter 3 describes reviews of the existing researches that are related to categorical data clustering using rough set theory.

Chapter 4 describes the proposed techniques for categorical data clustering, referred as Maximum Dependency of Attributes (MDA) technique. The notion of dependency of attributes in an information system using rough set theory, the correctness proof that the highest degree of dependency is the best clustering attribute selection are

presented. The accuracy measurement and the complexity of the technique also described. Finally, a technique for object splitting (partitioning) using divide and conquer technique and clusters validation for clusters purity measurement are presented.

Chapter 5 describes the experimental results of the proposed techniques. Empirical studies based on four small datasets, thirteen benchmark datasets and real world datasets demonstrate how the proposed technique performs better as compared with the rough set-based techniques. Further, an application of the proposed technique for clustering supplier chain management is presented. Discussion and analysis of the results of the proposed technique will be in detail here.

Finally, the conclusion and future work will be described in Chapter 6.



PTTA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH

## CHAPTER II

### 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.

## 2.1 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 2.1).

Table 2.1: An information system

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

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 2.2.

**Example 2.1.** Suppose that data about 6 students is given, as shown in Table 2.2.

Table 2.2: A student decision system

Student	Analysis	Algebra	Statistics	Decision
1	bad	good	medium	accept
2	good	bad	medium	accept
3	good	good	good	accept
4	bad	good	bad	reject
5	good	bad	medium	reject
6	bad	good	good	accept

The following values are obtained from Table 2.2,

$$U = \{1,2,3,4,5,6\},$$

$$A = \{\text{Analysis, Algebra, Statistics, Decision}\}, \text{ where}$$

$$C = \{\text{Analysis, Algebra, Statistics}\}, D = \{\text{Decision}\}$$

$$V_{\text{Analysis}} = \{\text{bad, good}\},$$

$$V_{\text{Algebra}} = \{\text{bad, good}\},$$

$$V_{\text{Statistics}} = \{\text{bad, medium, good}\},$$

$$V_{\text{Decision}} = \{\text{accept, reject}\}.$$

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

$f(u, a): U \times A \rightarrow V$  is a tuple  $t_i = (f(u_i, a_1), f(u_i, a_2), f(u_i, a_3), \dots, f(u_i, a_{|A|}))$ , 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 (refers to students 2 and 5 in Table 2.2). 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.

## 2.2 Indiscernibility Relation

From Table 2.2, it is noted that students 2, 3 and 5 are indiscernible (or similar or indistinguishable) with respect to the attribute Analysis. Meanwhile, students 3 and 6 are indiscernible with respect to attributes Algebra and Decision, and students 2 and 5 are indiscernible with respect to attributes Analysis, Algebra and Statistics. 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.

### 2.3 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.

## 2.4 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

- 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  $B$ ).
- 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

- 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  $B$ ).
- 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  $B$ .
- 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 2.1.

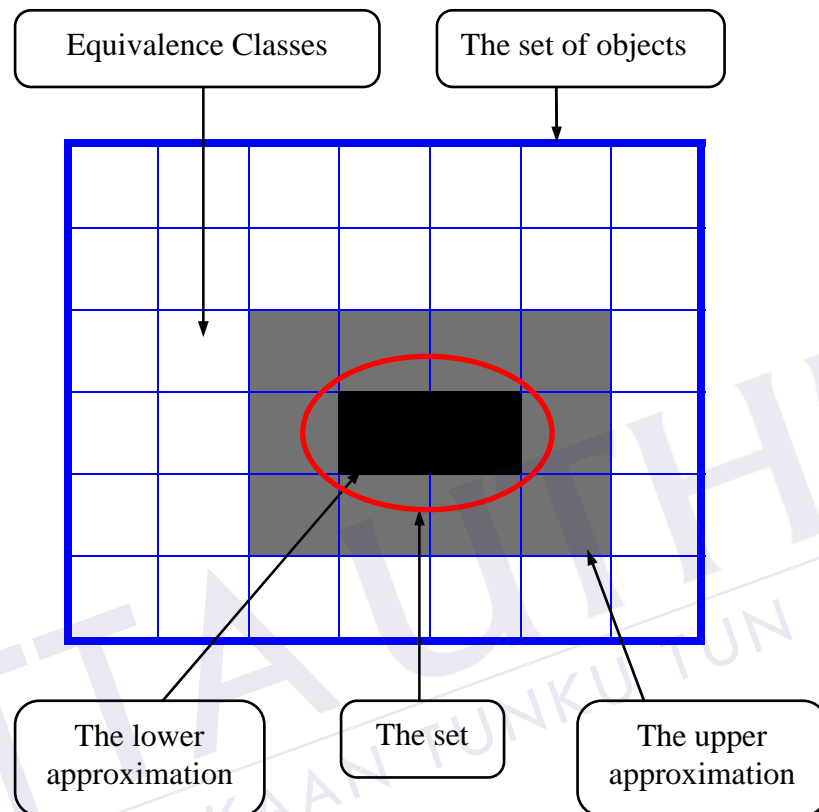


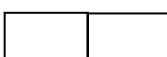


Figure 2.1: Set approximations

From Figure 2.1, three disjoint regions are given as follows

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

Let  $\phi$  be the empty set,  $X, Y \subseteq U$  and  $\neg X$  be the complement of  $X$  in  $U$ .

The lower and upper approximations satisfy the following properties (Zhu, 2007):

- |       |  |                             |
|-------|--|-----------------------------|
| (1L)  | $\underline{B}(U) = U$   | (Co-Normality)              |
| (1U)  | $\overline{B}(U) = U$  | (Co-Normality)              |
| (2L)  | $\underline{B}(\phi) = \phi$   | (Normality)                 |
| (2U)  | $\overline{B}(\phi) = \phi$  | (Normality)                 |
| (3L)  | $\underline{B}(X) \subseteq X$   | (Contraction)               |
| (3U)  | $X \subseteq \overline{B}(X)$  | (Extension)                 |
| (4L)  | $\underline{B}(X \cap Y) = \underline{B}(X) \cap \underline{B}(Y)$         | (Multiplication)            |
| (4U)  | $\overline{B}(X \cup Y) = \overline{B}(X) \cup \overline{B}(Y)$            | (Addition)                  |
| (5L)  | $\underline{B}(X \cup Y) \supseteq \underline{B}(X) \cup \underline{B}(Y)$ | (Inclusion)                 |
| (5U)  | $\overline{B}(X \cap Y) \subseteq \overline{B}(X) \cap \overline{B}(Y)$    | (Inclusion)                 |
| (6L)  | $\underline{B}(\underline{B}(X)) = \underline{B}(X)$                       | (Idempotency)               |
| (6U)  | $\overline{B}(\overline{B}(X)) = \overline{B}(X)$                          | (Idempotency)               |
| (7L)  | $\underline{B}(\neg X) = \neg \overline{B}(X)$                             | (Duality)                   |
| (7U)  | $\overline{B}(\neg X) = \neg \underline{B}(X)$                             | (Duality)                   |
| (8L)  | $X \subseteq Y \Rightarrow \underline{B}(X) \subseteq \underline{B}(Y)$    | (Monotone)                  |
| (8U)  | $X \subseteq Y \Rightarrow \overline{B}(X) \subseteq \overline{B}(Y)$      | (Monotone)                  |
| (9L)  | $\underline{B}(\neg \underline{B}(X)) = \neg \underline{B}(X)$             | (Lower Complement Relation) |
| (9U)  | $\overline{B}(\neg \overline{B}(X)) = \neg \overline{B}(X)$                | (Upper Complement Relation) |
| (10L) | $\forall G \in U / B, \underline{B}(G) = G$                                | (Granularity)               |
| (10U) | $\forall G \in U / B, \overline{B}(G) = G$                                 | (Granularity)               |

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. In (Herawan and Mat Deris, 2009e), it is shown that the property of (5L) and (5U) are properly inclusion.

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

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

where  $|X|$  denotes the cardinality of  $X$ . For empty set  $\phi$ , 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 2.2.

Consider the concept “Decision”, i.e., the set  $X$  (Decision = accept) =  $\{1,2,3,6\}$  and the set of attributes  $C = \{\text{Analysis, Algebra, Statistics}\}$ . The partition of  $U$  induced by  $IND(C)$  is given by

$$U / C = \{\{1\}, \{2,5\}, \{3\}, \{4\}, \{6\}\}.$$

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

$$\underline{C}(X) = \{1,3,6\} \text{ and } \overline{C}(X) = \{1,2,3,5,6\}.$$

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

$$\alpha_C(X) = \frac{3}{5}.$$

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.

## 2.5 Summary

In this chapter, the concept of rough set theory through data contained in an information system has been presented. The rough set approach seems to be of fundamental importance to artificial intelligent, especially in the areas of decision analysis and knowledge discovery from databases (Pawlak, 1997; Pawlak, 2002; Pawlak, 2002). Basic ideas of rough set theory and its extensions, as well as many interesting applications can be found in (Peters and Skowron, 2006; <http://roughsets.home.pl/www/>; [http://en.wikipedia.org/wiki/Rough\\_set](http://en.wikipedia.org/wiki/Rough_set)).

## CHAPTER III

### CATEGORICAL DATA CLUSTERING USING ROUGH SET THEORY

This chapter describes and review related existing researches in categorical data clustering using rough set theory. It also includes the advantages and disadvantages of recent works that have been done in this field.

#### 3.1 Data Clustering

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters (Guha, Rastogi and Shim, 1998). Clustering may be found under different names in different contexts, such as unsupervised learning, numerical taxonomy, typology and partition.

In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data that is why it is perceived as an unsupervised process. On the other hand, classification is a procedure of assigning a data item to a predefined set of categories (Piatesky-Shapiro,



Fayyad and Smyth, 1996). Clustering produces initial categories in which values of a data set are classified during the classification process. The clustering process may result in different partitioning of a data set, depending on the specific criterion used for clustering. Thus, there is a need of preprocessing before a clustering task is assumed in a data set.

The basic steps to develop clustering process can be summarized as follows (Piatesky-Shapiro, Fayyad and Smyth, 1996):

- a. *Feature selection*. The preprocessing of data may be necessary prior to their utilization in clustering task.
- b. *Clustering algorithm*. This step refers to the choice of an algorithm that results in the definition of a good clustering scheme for a data set.
- c. *Validation of the results*. The correctness of clustering algorithm results is verified using appropriate criteria and techniques. Since clustering algorithms define clusters that are not known a priori, irrespective of the clustering methods, the final partition of data requires some kind of evaluation in most applications (Rezaee, Lelieveldt and Reiber, 1998).
- d. *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.

According to the method adopted to define clusters, clustering algorithms can be broadly classified into the following types (Jain, Murty and Flyn, 1999):

- a. *Partitional clustering*. It attempts to directly decompose the data set into a set of disjoint clusters.
- b. *Hierarchical clustering*. It proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters.
- c. *Density-based clustering*. The key idea of this type of clustering is to group neighboring objects of a data set into clusters based on density conditions.
- d. *Grid-based clustering*. This type of algorithms is mainly proposed for spatial data mining.

For each of above categories there is a wealth of subtypes and different algorithms for finding the clusters. Thus, according to the type of variables allowed

in the data set can be categorized into (Rezaee, Lelieveldt and Reiber, 1998; Guha, Rastogi and Shim, 2000; Huang, 1997):

- a. *Statistical*, which are based on statistical analysis concepts. They use similarity measures to partition objects and they are limited to numeric data.
- b. *Conceptual*, which are used to cluster categorical data. They cluster objects according to the concepts they carry. Another classification criterion is the way clustering handles uncertainty in terms of cluster overlapping.
- c. *Fuzzy clustering*, which uses fuzzy techniques to cluster data and they consider that an object can be classified to more than one clusters. The most important fuzzy clustering algorithm is *Fuzzy C-Means* (Bezdeck, Ehrlich and Full, 1984).
- d. *Crisp clustering*, considers non-overlapping partitions meaning that a data point either belongs to a class or not.
- e. *Kohonen net clustering*, which is based on the concepts of neural networks. The Kohonen network has input and output nodes.

In general terms, the clustering algorithms are based on a criterion for assessing the quality of a given partitioning. More specifically, they take as input some parameters (e.g. number of clusters, density of clusters) and attempt to define the best partitioning of a data set for the given parameters. Thus, they define a partitioning of a data set based on certain assumptions and not necessarily the “best” one that fits the data set (Halkidi, Batistakis and Vazirgiannis, 2001).

### 3.2 Categorical Data Clustering

Nowadays, much of the data in databases is categorical: fields in tables whose attributes cannot naturally be ordered as numerical values can. As a concrete example; consider a database describing car sales with attributes manufacturer model, dealer, price, color, customer and sale date. In our view, price and sale date are traditional numerical values. Color is arguably a categorical value, assuming that values such as *red* and *green* cannot easily be ordered linearly. Attributes such as manufacturer model and dealer are indisputably categorical attributes. And it is very

hard to reason that one dealer is *like* or *unlike* another in the way one can reason about numbers.

Several works concerning on categorical data clustering have been proposed. Ralambondrainy proposed a method to convert multiple categorical attributes into binary attributes using 0 and 1 to represent either a category absence or presence (Ralambondrainy, 1995). Ganti *et al.* proposed CACTUS (Clustering Categorical Data Using Summaries), a summarization based algorithm (Ganti, Gehrke and Ramakrishnan, 1999). In CACTUS, the authors cluster for categorical data by generalizing the definition of a cluster for numerical attributes. Summary information constructed from the data set is assumed to be sufficient for discovering well-defined clusters. CACTUS finds clusters in subsets of all attributes and thus performs a subspace clustering of the data. Guha *et al.* proposed a hierarchical clustering method termed ROCK (Robust Clustering using Links), which can measure the similarity or proximity between a pair of objects (Guha, Rastogi and Shim, 2000). Using ROCK, the number of ‘links’ are computed as the number of common neighbors between two objects. An agglomerative hierarchical clustering algorithm is then applied: first, the algorithm assigns each object to a separate cluster, clusters are then merged repeatedly according to the closeness between clusters, where the closeness is defined as the sum of the number of ‘links’ between all pairs of objects. Zaki *et al.* proposed a novel algorithm for mining subspace clusters in categorical datasets called CLICKS (Zaki, Peters, Assent and Seidl, 2007). The CLICKS algorithm finds clusters in categorical datasets based on a search for k-partite maximal cliques. Unlike CACTUS and ROCK, CLICKS mines subspace clusters. The results confirmed that CLICKS is superior to both CACTUS and ROCK in detecting even the simplest of clusters and the faster clustering process, respectively. Some of the methods mentioned above, such as CACTUS and ROCK algorithms have one common assumption: each object can be classified into only one cluster and all objects have the same degree of confidence when grouped into a cluster. However, in real world applications, it is difficult to draw sharp boundaries between the clusters. Therefore, the uncertainty of the objects belonging to the cluster needs to be considered.

Many theories, techniques and algorithms have been developed to deal with the problem of uncertainty. In 1960’s, Zadeh proposed completely new, elegant approach to handle uncertainty called fuzzy sets theory (FST). Since fuzzy sets has

been proposed, fuzzy-based clustering has been widely studied and applied in a variety of substantive areas. For categorical data clustering, Kim *et al.* use fuzzy centroids to represent the clusters of categorical data instead of the hard-type centroids (Kim, Lee and Lee, 2004). The use of fuzzy centroids makes it possible to fully exploit the power of fuzzy sets in representing the uncertainty in the classification of categorical data. For rough set theory, Chen *et al.* proposed a technique called Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data (RAHCA). Nevertheless, the clustering problem in RAHCA is described using the clustering decision table. Thus, the decision attribute must be given in order to do clustering using RAHCA (Chen, Cui, Wang and Wang, 2006).

The problem of clustering categorical data involves complexity not encountered in the corresponding problem for numerical data, since one has much less a priori structure to work with. Clustering techniques for categorical data are very different from those for numerical data in terms of the definition of similarity measure. Traditionally, categorical data clustering is merged into numerical clustering through a data preprocessing stage (Jain, Murty and Flynn, 1999). In the feature selection (preprocessing) process, numerical features are constructed from the categorical data, or a conceptual similarity function between data records is defined based on the domain knowledge. However, meaningful numerical features or conceptual similarity are usually difficult to extract at the early stage of data analysis, because one has little knowledge about the data. It has been widely recognized that directly clustering the raw categorical data is important for many applications (Chen and Liu, 2009), without any pre-processing process. Therefore, there are increasing interests in clustering raw categorical data (Parmar, Wu and Blackhurst, 2007; Huang, 1997; Huang, 1998; Kim, Lee and Lee, 2004).

### **3.3 Categorical Data Clustering using Rough Set Theory**

In dealing with the raw categorical data, one of the difficulties is to resolve the problem of similarity measure, for the nature of categorical data is the non-numerical so that Euclidean distance extensively-used in numerical data processing can not be employed directly. However, rough set theory can be applied to clustering analysis;

## REFERENCES

- Agrawal, R., Imielinski, T., and Swami, A. (1993). *Database Mining: A Performance Perspective*. IEEE Transactions on Knowledge and Data Engineering, 5 (6), 914–925.
- Atkinson-Abutridy, J., Mellish, C., and Aitken, S. (2004). *Combining information extraction with genetic algorithms for text mining*. IEEE Intelligent Systems, 19 (3), 22–30.
- Bezdeck, J.C., Ehrlich, R., and Full, W. (1984). *FCM: Fuzzy C-Means Algorithm*. Computers and Geoscience, 10 (2–3), 191–203.
- Bi, Y., Anderson, T. and McClean, S. (2003). *A rough set model with ontologies for discovering maximal association rules in document collections*. Knowledge Based Systems, 16, 243–251.
- Busse, J.G. (1998). *Knowledge Discovery under Uncertainty: A rough set Approach*. Journal of Intelligent Robotics System, 1, 3–16.
- Chen, D., Cui, D., Wang, C. and Wang, Z. (2006). *Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data*. International Journal of Information Technology, 12 (3), 149–159.
- Chen, K. and Liu, L. (2009). *“Best K”: critical clustering structures in categorical datasets*. Knowledge and Information System, 20, 1–33.

- Chouchoulas, A and Shen, Q. (2001). *Rough set-aided keyword reduction for text categorization*. Applied Artificial Intelligence, 15 (9), 843–873.
- Dempster, A., Laird, N., and Rubin, D. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society 39 (1), 1–38.
- Duntsch, I., and Gediga, G. (2000). *Rough Set Data Analysis: A road to non-invasive knowledge discovery*. Bangor: Methodos.
- Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). *CACTUS – clustering categorical data using summaries*. In Proceeding of Fifth ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, 73–83.
- Gibson, D., Kleinberg, J. and Raghavan, P. (2000). *Clustering categorical data: an approach based on dynamical systems*. The Very Large Data Bases Journal 8 (3–4) 222–236.
- Guan, J.W., Bell, D.A. and Liu, D.Y. (2003). *The Rough Set Approach to Association Rule Mining*. In the Proceedings of the Third IEEE ICDM'03, 529–532.
- Guan, J.W., Bell, D.A. and Liu, D.Y. (2005). *Mining Association Rules with Rough Sets*. Studies in Computational Intelligence, Springer Verlag, 163–184.
- Guha, S., Rastogi, R., and Shim K. (1998). *CURE: An Efficient Clustering Algorithm for Large Databases*. In the Proceeding of the ACM SIGMOD Conference'98.
- Guha, S., Rastogi, R. and Shim, K. (2000). *ROCK: a robust clustering algorithm for categorical attributes*. Information Systems 25 (5), 345–366.

- Haimov, S., Michalev, M., Savchenko, A., and Yordanov, O. (1989) *Classification of radar signatures by autoregressive model fitting and cluster analysis*. IEEE Transactions on Geo Science and Remote Sensing 8 (1), 606–610.
- Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). *On clustering validation techniques*. Journal of Intelligent Information Systems 17 (2–3) 107–145.
- Herawan, T. and Mat Deris, M. (2009a). *Rough set theory for topological space in information systems*. In the Proceeding of International Conference of AMS 2009, IEEE Press, 107–112.
- Herawan, T. and Mat Deris, M. (2009b). *A direct proof of every rough set is a soft set*. In the Proceeding of International Conference of AMS 2009, IEEE Press, 119–124.
- Herawan, T. and Mat Deris, M. (2009c). *A construction of nested rough set approximation in information systems using dependency of attributes*. In the Proceeding of International Conference of PCO 2009, American Institute of Physic 1159, 324–331.
- Herawan, T. and Mat Deris, M. (2009d). *Rough set theory for selecting clustering attribute*. In the Proceeding of International Conference of PCO 2009, American Institute of Physic 1159, 331–338.
- Herawan, T. and Mat Deris, M. (2009e). *Rough topological properties of set in information systems using dependency of attributes*. In the Proceeding of International Conference of CITA 2009, 39–46.
- Herawan, T. and Mat Deris, M. (2009f). *A framework on rough set based partitioning attribute selection*. Lecture Notes in Artificial Intelligence 5755 (1), Springer Verlag, 91–100.

- Herawan, T. and Mat Deris, M. (2009h). *On multi-soft sets construction in information systems*. Lecture Notes in Artificial Intelligence 5755 (1), Springer Verlag, 101–110.
- Herawan, T., Mohd Rose, A.N. and Mat Deris, M. (2009). *Soft set theoretic approach for dimensionality reduction*. Communication of Computer and Information Sciences 64, Springer-Verlag, 180–187.
- Herawan, T., Yanto, I.T.R. and Mat Deris, M. (2009a). *SMARViz: Soft maximal association rules visualization*. Lecture Notes in Computer Science 5857, Springer Verlag, 664–674.
- Herawan, T., Yanto, I.T.R. and Mat Deris, M. (2009b). *A soft set approach for maximal association rules mining*. Communication of Computer and Information Sciences 64, Springer-Verlag, 163–170.
- Herawan, T., Yanto, I.T.R. and Mat Deris, M. (2009c). *Rough set approach for categorical data clustering*. Communication of Computer and Information Sciences 64, Springer Verlag, 188–195.
- Herawan, T., Ghazali, R. and Mat Deris, M. (2010). Soft set theoretic approach for dimensionality reduction. *International Journal of Database Theory and Application* 3 (1), 47–60.
- Herawan, T., Ghazali, R., Yanto, I.T.R. and Mat Deris, M. (2010). Rough set approach for categorical data clustering. Manuscript to appear in *International Journal of Database Theory and Application* 3 (1), 33–52.
- Herawan, T., Yanto, I.T.R. and Mat Deris, M. (2010a). *A construction of hierarchical rough set approximations in information systems using dependency of attributes*. Studies in Computational Intelligence, Springer Verlag, 3–15.



Herawan, T., Yanto I.T.R. and Mat Deris, M. (2010b). ROSMAN: ROugh Set approach for clustering supplier chain MANagement, Manuscript accepted in a special issue of Soft Computing Methodology, *International Journal of Biomedical and Human Sciences*, Japan, to appear in Vol. 17, No. 1, July 2010.

Herawan, T., Mat Deris, M. and Abawajy, J.H. (2010a). *Rough set approach for selecting clustering attribute*. Knowledge Based Systems, Elsevier, 23 (3), 220–231.

Herawan, T., Mat Deris, M. and Abawajy, J.H. (2010b). *Matrices representation of multi soft-sets and its application*. Lecture Notes in Computer Science, Springer Verlag, to appear in D. Taniar et al. (Eds.): ICCSA 2010, Part III, LNCS 6018, 201–214, 2010.

Herawan, T. and Mat Deris, M. (2010a). *Soft decision making for patients suspected influenza*. Lecture Notes in Computer Science, Springer Verlag, to appear in D. Taniar et al. (Eds.): ICCSA 2010, Part III, LNCS 6018, 405–418, 2010.

Herawan, T. and Mat Deris, M. (2010b). *Soft set theory for association rules mining*. Manuscript accepted in Knowledge Based Systems, Elsevier.

Herawan, T., Mohd Rose, A.N. and Mat Deris, M. (2010). *Soft set theoretic approach for discovering attributes dependency in information systems*. In L. Zhang, J. Kwok, and B.L. Lu (Eds.): ISNN 2010, Part II, Lecture Notes in Computer Science 6064, pp. 596–605, 2010. © Springer-Verlag Berlin Heidelberg 2010.

<http://archive.ics.uci.edu/ml/datasets/Soybean>

<http://archive.ics.uci.edu/ml/datasets/Zoo>

[http://en.wikipedia.org/wiki/Rough\\_set](http://en.wikipedia.org/wiki/Rough_set)

<http://roughsets.home.pl/www/>

<http://www.ics.uci.edu/~mlearn/MLRepository.html>

Hu, X. (1995). *Knowledge discovery in databases: An attribute oriented rough set approach*. PhD Thesis, Department of Computer Science, University of Regina, Canada.

Huang, Z. (1997). *A Fast Clustering Algorithm to Cluster very Large Categorical Data Sets in Data Mining*. Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97), in cooperation with ACM SIGMOD'97.

Huang, Z. (1998). *Extensions to the k-means algorithm for clustering large data sets with categorical values*. *Data Mining and Knowledge Discovery* 2 (3) 283–304.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). *Data Clustering: A Review*. *ACM Computing Surveys*, 31 (3), 264–323.

Jensen, R. (2005). *Combining rough and fuzzy sets for feature selection*. Ph.D. Thesis, School of Informatics University of Edinburgh.

Jiang, D., Tang, C., and Zhang, A. (2004). *Cluster analysis for gene expression data: a survey*. *IEEE Transactions on Knowledge and Data Engineering* 16 (11), 1370–1386.

Kim, D., Lee, K. and Lee, D. (2004). *Fuzzy clustering of categorical data using fuzzy centroids*. *Pattern Recognition Letters* 25 (11) 1263–1271.

Komorowski, J., Polkowski, L. and Skowron, A. (1999). *Rough sets: a tutorial*, in S.K. Pal and A. Skowron, editors, *Rough-Fuzzy Hybridization*, Springer-Verlag, 3–98.

Langley, P., Iba, W., and Thompson, K. (1992). *An Analysis of Bayesian Classifiers*. In *Proceeding of the Tenth National Conference on Artificial Intelligence*, 223–228.

Magnani, M. (2005). *Rough Set Theory for Knowledge Discovery in Data Bases*. Technical Report, Department of Computer Science, University of Bologna, Italy, 1–15.

Mazlack, L.J. (1996). *Database Mining by Learned Cognitive Dissonance Reduction*. Proceedings of the Fifth IEEE International Conference of Fuzzy Systems (FUZZ-IEEE'96).

Mazlack, L.J., He, A., Zhu, Y. and Coppock, S. (2000). *A rough set approach in choosing clustering attributes*. In the Proceedings of the ISCA 13th, International Conference CAINE-2000, 1–6.

Mathieu, R., and Gibson, J. (2004). *A Methodology for large scale R&D planning based on cluster analysis*. IEEE Transactions on Engineering Management 40 (3), 283–292.

Mohd Rose, A.N., Herawan, T. and Mat Deris, M. (2010). *A framework of decision making based on maximal supported sets*. In L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISSN 2010, Part I, Lecture Notes in Computer Science 6063, pp. 473–482, 2010. © Springer-Verlag Berlin Heidelberg 2010.

Molodtsov, D. (1999). *Soft set theory-first results*. Computers and Mathematics with Applications. 37, 19–31.

Parmar, D., Wu, T. and Blackhurst, J. (2007). *MMR: An algorithm for clustering categorical data using rough set theory*. Data and Knowledge Engineering 63, 879–893.

- Parmar, D., Wu, T., Callarman, T., Fowler, J. and Wolfe, P. (2009). *A Clustering Algorithm for Supplier Base Management*. International Journal of Production Research, DOI: 10.1080/00207540902942891.
- Pawlak, Z. (1982). *Rough sets*. International Journal of Computer and Information Science, 11, 341–356.
- Pawlak, Z. (1983). *Rough classification*. International Journal of Human Computer Studies 51, 369–383.
- Pawlak, Z. (1985). *Rough set and Fuzzy sets*. Fuzzy sets and systems. 17, 99–102.
- Pawlak, Z. (1991). *Rough sets: A theoretical aspect of reasoning about data*. Kluwer Academic Publisher.
- Pawlak, Z. (1997). *Rough set approach to knowledge-based decision support*. European Journal of Operational Research, 99, 48–57.
- Pawlak, Z. (2002). *Rough sets and intelligent data analysis*. Information Sciences, 147, 1–12.
- Pawlak, Z. (2002). *Rough set, decision algorithm and Bayes's theorem*. European Journal of Operational Research, 136, 181–189
- Pawlak, Z. and Skowron, A. (2007). *Rudiments of rough sets*. Information Sciences 177 (1), 3–27.
- Pawlak, Z. and Skowron, A. (2007). *Rough sets: Some extensions*. Information Sciences 177 (1), 28–40.
- Peters, J.F. and Skowron, A. (2006). *Zdzislaw Pawlak: life and work (1926-2006)*. In J.F. Peters and A. Skowron, editors, Transaction on Rough Set V, LNCS 4100, Springer-Verlag, 1–24.

- Piatessky-Shapiro G., Fayyad, U. and Smyth, P. (1996). *From data mining to knowledge discovery: an overview*. Advances in Knowledge discovery and Data Mining, AAAI/MIT Press, 1–34.
- Quinlan, J.R. (1993). *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann.
- Ralambondrainy, H. (1995). *A conceptual version of the K-means algorithm*. Pattern Recognition Letters 16 (11), 1147–1157.
- Rezaee, R., Lelieveldt, B.P.F., and Reiber, J.H.C. (1998). *A New Cluster Validity Index for the Fuzzy c-Mean*. Pattern Recognition Letters, 19, 237–246.
- Roiger, R.J. and Geatz, M.W. (2003). *Data Mining: A Tutorial-Based Primer*. Addison Wesley.
- Sever, H. (1998). *The status of research on rough sets for knowledge discovery in databases*. In Proceedings of the Second International Conference on Nonlinear Problems in Aviation and Aerospace (ICNPAA98), 2, 673–680.
- Shen, W. (1991). *Discovering Regularities from Large Knowledge Bases*. In Proceeding of Eighth International Workshop of Machine Learning, 1991, 539–543.
- Smyth, P., and Goodman, R. (1992). *An Information Theoretical Approach to Rule Induction From Databases*. IEEE Transactions on Knowledge and Data Engineering, 4 (4), 85–97.
- Thomas, D.J. and Griffin, P.M. (1996). *Coordinated supply chain management*. European Journal of Operational Research, 94, 1–15.
- Wong, K., Feng, D., Meikle, S., and Fulham, M. (2002). *Segmentation of dynamic pet images using cluster analysis*. IEEE Transactions on Nuclear Science 49 (1), 200–207.

- Wu, S., Liew, A., Yan, H., and Yang, M. (2004). *Cluster analysis of gene expression data based on self-splitting and merging competitive learning*. IEEE Transactions on Information Technology in BioMedicine 8 (1), 5–15.
- Yanto, I.T.R., Herawan, T. and Mat Deris, M. (2010a). *Data clustering using VPRS*. Manuscript accepted to appear in *Intelligent Data Analysis*, IOS Press.
- Yanto, I.T.R., Herawan, T. and Mat Deris, M. (2010b). *A framework on rough set approach for clustering web transaction*. Studies in Computational Intelligence, Springer Verlag, 265–277.
- Yanto, I.T.R., Herawan, T. and Mat Deris, M. (2010c). RoCeT: Rough set approach for clustering web transactions, Manuscript accepted in a special issue of Soft Computing Methodology, *International Journal of Biomedical and Human Sciences*, to appear in Vol. 17, No. 1, 2010.
- Yao, Y.Y. (1996). *Two views of the theory of rough sets in finite universes*. Approximate Reasoning 15 (4), 191–317.
- Yao, Y.Y. (1998). *Constructive and algebraic methods of the theory of rough sets*. Information Sciences 109 (1–4), 21–47.
- Yao, Y.Y. (1998). *Relational interpretations of neighborhood operators and rough set approximation operators*. Information Sciences 111, 239–259.
- Yao, Y.Y. (2001). *Information granulation and rough set approximation*. International Journal of Intelligent Systems 16 (1), 87–104.
- Zadeh, L.A. (1965). *Fuzzy sets*. Information and Control, 8, 338–353.
- Zaki, M.J., Peters, M., Assent, I., and Seidl, T. (2007). *Clicks: An effective algorithm for mining subspace clusters in categorical datasets*. Data and Knowledge Engineering, 60 (1), 51–70.

Zhu, W. (2007). *Topological approaches to covering rough sets*. Information Sciences 177, 1499–1508.

Ziarko, W. (1991). *Variable precision rough set model*. Journal of computer and system science 46, 39–59.

