Chapter 1　Introduction

1.1 Overview

Traditionally, analysts have performed the task of extracting useful information from recorded data. But, the increasing volume of data in modern business and science calls for computer-based approaches. As data sets have grown in size and complexity, there has been an inevitable shift away from direct hand on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools. The modern technologies of computers, networks and sensors have made data collection and organization an almost effortless task. However, the gathered data needs to be converted into information and knowledge from recorded data to become useful. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery from data (Kantardzic, 2003).

Knowledge bases have been successfully applied in many real world applications, where intelligent decisions have to be made. Knowledge bases can usually be represented as a set of decision rules that generally follow the form of:　"if…, then…". These rules can be extracted from human experts or collected data. Most of the time the collected data is so huge that it is beyond the ability of a human expert to analyze it without using feasible analysis techniques. The analysis and extraction of useful information from collected data has been the subject of active research in data mining (Lin, et al.1997).

Fundamental issues in knowledge discovery arise from the very nature of databases and the objects they deal with. They are characterized as follows (Cios, et al. 1998):

(1) Huge volume of data:

Many data mining techniques are very sensitive to the size of data in terms of time complexity. Thus it becomes an imperative that the searches over large spaces of possible relations between attributes values and classes must use heuristics or reduce the search space either attributes or objects.

(2) Incomplete or imprecise data:

In many practical situations the information collected in a database can be either incomplete or imprecise.

(3) Noisy data:

It is very difficult during data collection to eliminate or just reduce the amount of non-systematic errors, call noise. This is a serious problem that implies that data mining methods should be made less sensitive to noise.

Rough sets as a mathematical methodology for data analysis were introduced by Pawlak (Pawlak, 1991). They provide a powerful tool for data analysis and knowledge discovery from imprecise and ambiguous data. The theory of rough sets has been successfully applied to diverse areas, such as knowledge acquisition, forecasting and predictive modeling, knowledge base systems, and data mining (Slowinski, 1992; Beynon, et al. 2001). The traditional reduct generated by the rough sets approach, the reduct was employed for reduces redundant attributes as well as redundant objects from the decision table. Therefore, the reduct contains less "noisy" data and provides a decision table that can yield a substantially lower misclassification rate (Hashemi, et al. 1998).

The rough sets methodology is based on the premise that lowering the degree of precision in the data makes the data pattern more visible. The central premise of the rough sets philosophy is that knowledge exists in the ability to classify. In other words, the rough sets approach can be considered as a formal framework for discovering patterns from imperfect data. The results of the rough sets approach are presented in the form of classification or decision rules derived from given data sets.

The rough sets approach is inspired by the notion of inadequacy of the available information to perform a complete classification of objects; that is, performing a complete classification requires that the collected data must be fully correct or certain. The classification with a controlled degree of uncertainty, or a misclassification error, is outside the realm of this approach (Ziarko, 1993). The variable precision rough sets (VPRS) model was introduced by Ziarko (1993) and is an extension of the original Rough Set Theory (RST) as a tool for classification of objects. This is an important extension, since as noted by Kattan, et al. (1998), 'In real world decision making, the patterns of classes often overlap, suggesting that predictor information may be incomplete… This lack of information results in probabilistic decision making, where perfect prediction accuracy is not expected.

1.2 Research Motivation

The Variable Precision Rough Sets (VPRS) model is a powerful tool for data mining, as it has been widely applied to acquire knowledge. Despite its diverse applications in many domains, the VPRS model unfortunately cannot be applied to real world classification tasks involving continuous attributes. This requires a discretization method to pre-process the data. Discretization is an effective technique in dealing with continuous attributes for data mining, especially for the classification problem. Many classification algorithms require that the training data contain only discrete attributes, and some would work better on discretized or binarized data (Li, et al. 2002; Kerber, 1992). However, for these algorithms, discretizing continuous attributes is a first step for deriving classification rules.

VPRS deals with partial classification by introducing a precision parameter β (in the rough set the β value is zero). Ziarko (1993) considered β as a classification error and it is defined to be in the domain $[0.0, 0.5)$. However, VPRS lacks a feasible approach to determine β. Without a β value to control the selection of β-reducts, this may lead to the full set of β-reducts becoming too large, such that an addition to a search scope is needed to find a suitable β-reduct.

1.3 Research Objectives

The VPRS procedure for data mining is shown in figure 1.1. This research aims to propose an effective disctetization algorithm for VPRS model, and we will compare the performance of our proposed algorithm with other RS-based discretization algorithms. In addition, we propose a two-step approach to select the β-reducts. A comparison of the performance between the proposed approach and other β-reducts selection methods will be made.

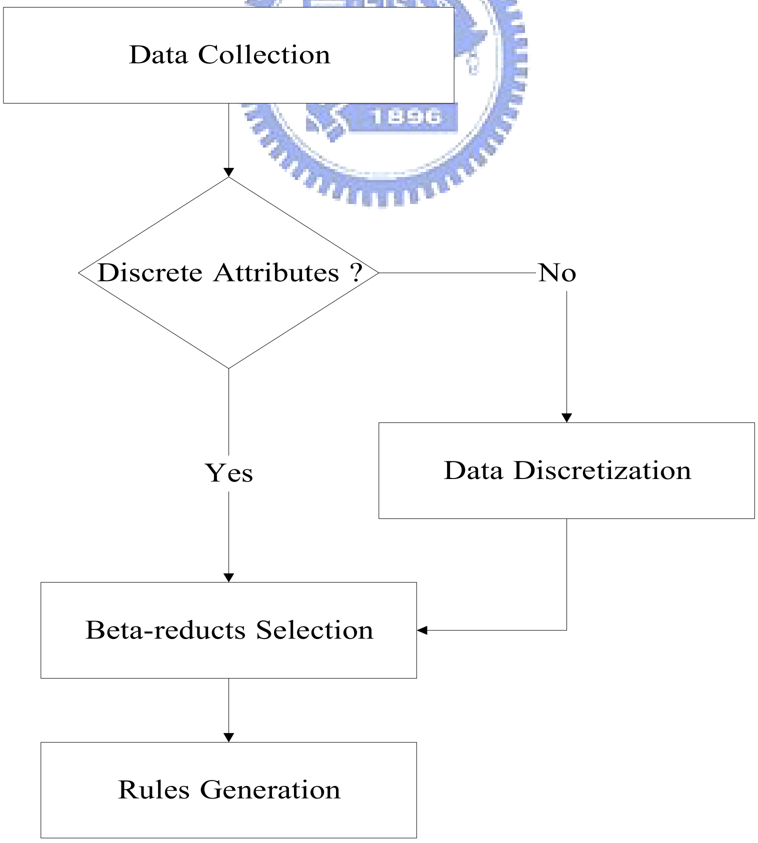

Figure 1.1 VPRS Procedure for Data Mining

1.4 Organization

The organization of remaining chapters for this study is as follows: Chapter 2 reviews the literature of discretization, the β-reducts and describes the original rough sets theory and variable precision rough sets model. Moreover, we also briefly review the relation rules extraction methods. Chapter 3 introduces the variable precision rough sets theory. Chapter 4 presents the proposed discretization algorithm, named the extended Chi2 algorithm and an approach for the β-reducts. Chapter 5 illustrates the extended Chi2 algorithm through five data sets and select β-reducts approach through a real medical examination case. In Chapter 6, we show that the VPRS theory using proposed procedure can be applied to a communication industrial case. Finally, the conclusions and the direction of further research are given in Chapter 7.