SOFT SET APPROACH FOR DECISION ATTRIBUTE SELECTION

IN DATA CLUSTERING

LOK LEH LEONG

THESIS SUBMITTED IN FULFILMENT OF THE DEGREE OF COMPUTER SCIENCE (SOFTWARE ENGINEERING)

FACULTY OF COMPUTER SYSTEM AND SOFTWARE ENGINEERING

UNIVERSITI MALAYSIA PAHANG

2013

ABSTRACT

Clustering is one of the fundamental operations in data mining that cluster set of heterogeneous data objects into smaller homogeneous classes. Using clustering attribute (decision attribute) is one of the data clustering techniques. Soft set theory is a new mathematical tool applying in clustering applications in databases circumstances. Hence, the research aim is to find the practical technique of soft set theory for decision attribute selection in soft set theory. The test is been done by using two UCI benchmark datasets to determine the speed of execution time for soft set approach with rough set techniques, that are Total Roughness (TR), Min-Min Roughness (MMR) and Maximum Dependency of Attributes (MDA). The results show that the proposed technique provides faster decision for selecting a clustering attribute.

ABSTRAK

Menyusun dalam kelompok adalah salah satu operasi asas dalam perlombongan data untuk menyusun dari kumpulan yang berlainan jadi kumpulan kecil yang mempunyai persamaan. Dengan menggunakan sifat kelompok adalah salah satu teknik untuk kelompok data. Teori set lembut adalah alat matematik yang baru untuk kelompok data dengan mengikut keadaan pangkalan data. Jadi, tujuan kajian adalah mahu mencari teknik teori set lembut yang praktikal dalam mencari sifat kelompok. Ujian ini akan dijalankan dengan menggunakan dua UCI set data untuk menentukan masa yang diperlukan untuk mencari sifat kelompok dengan membandingkan teori set lembut dan teori set kasar, iaitu Total Roughness (TR), Min-Min Roughness (MMR) dan Maximum Dependency of Attributes (MDA). Keputusan menunjukan bahawa teori yang dicadangkan adalah lebih laju dalam memilih sifat kelompok.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TR	-	Total Roughness
MMR	-	Min-Min Roughness
MDA	-	Maximum Dependency Attributes
DM	-	Data Mining
MZ	-	Marczeweski-Steinhaus metric

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

Chapter I will slightly explain on the general view of this research. This chapter divides into five parts which the first section is introduction. Next will continues by the problem statement. Then are the objectives which stated the goals to be achieved in this research. After that are the scopes of the system where the research's boundary is determined and finally is thesis organization which explain the summary of each chapter for this research.

.

1.1 INTRODUCTION

Clustering objects set into classes of homogeneous is a data mining fundamental operation. The operation is required in a number of data analysis tasks, like data summation and unsupervised classification, as well as segmentation of large homogeneous data sets into smaller homogeneous subsets that can be easily modeled separated, managed and analyzed. The clustering dataset is mapped as the decision table is one of the data clustering main idea that can be done by introducing a decision attribute. However, one practical problem is faced: for many candidates that exist in database, we need to select only one that is the best attribute to partition the objects.

Currently, there have been works in area that applying rough set theory in the process of selecting clustering attribute. Mazlack et al.proposed two techniques to select clustering attribute: Bi-Clustering (BC) technique based on balanced (unbalanced) 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. Palmar et al proposed a new technique called Min-Min Roughness (MMR) for categorical data clustering. In selecting clustering attribute, MMR is proposed to improve Mazlack's technique for datasets with multi-valued attributes. In order to improve the accuracy and computational complexity, the technique maximum dependency of attributes (MDA) is proposed. It is based on the dependency of attributes using rough set theory in an information system. However, since the algorithm for categorical data clustering based on rough set theory is relatively new, the focus of MMR algorithm has been on evaluate the performance. In reviewing TR, MMR and MDA techniques for handling large datasets, the ever-increasing computing capabilities, computation complexity is still an outstanding issue. This is due to all attributes are considered to be selected.

The soft set theory proposed by Molodtsov 1999 is a new method for handling uncertain data. Soft sets are called elementary neighborhood systems. Molodtsov pointed out that one of the main advantages of soft set theory is that it is free from the inadequacy of the parameterization tools, like in the theories of fuzzy set, probability and interval mathematics. As for standard soft set, it may be redefined as the classification of objects in two distinct classes, thus confirming that soft set can deal with a Boolean-valued information system.

For a multi-valued information system, the concept of multi-soft sets proposed by will be used. Since every rough set can be considered as soft set, then by applying the concept of a mapping inclusion in soft set theory propose an alternative technique for selecting clustering attribute under soft set theory. The main purpose of the proposed technique is to ensure that the process of partitioning attribute selection is used in transforming an information system into a decision system.

1.2 PROBLEM STATEMENTS

The concept of soft set theory introduces by Molodtsov is uses in generic mathematical tools to solve uncertainty data. Research in soft set theory field has been growth dynamic and great progress has been achieved in this few years. This field area includes the works by using of theoretical soft set theory, soft set approach in forecasting, soft set theory in abstract algebra, data analysis by using soft set theory, particularly in area of decision making and parameterization reduction. However, currently not researches have been done by applying soft set theory for selecting decision attribute in data clustering. Enlighten by the fact that every rough set is a soft set, in this paper, we show the soft set theory applicability for discovering soft set technique to select clustering attribute on benchmark datasets. This proposed approach is based on the notion of multi-soft sets. In order to select decision attribute, an inclusion of value sets in soft set theory is used. The results obtained that our proposed technique is equivalent to the rough-set based decision attribute selection.

1.3 OBJECTIVES

The objectives of the research are as follows:

i. To propose soft set based technique to select clustering attribute.
ii. To validate the proposed technique on benchmark datasets.
iii. To compare the proposed technique with that other baseline technique such as TR, MMR and MDA.

1.4 SCOPES

The scopes of this project are:

i. The proposed technique using soft set theory.
ii. The proposed technique is only for selecting clustering attribute.

1.5 THESIS ORGANIZATION

This thesis consists of six chapters. The following chapter of this thesis is organized as follows: Chapter I will discuss the research background. Chapter II will describes the data mining and decision attribute selection. Chapter III discusses the fundamental concept of soft set theory and rough sets which consists of TR, MMR and MDA. Chapter IV will explains the design and implementation system by using soft set approach. Chapter V elaborates the result of using soft set approach in determining decision attribute and following by discussion. Last, the research conclusion will describe in Chapter VI.

CHAPTER 1I

LITERATURE REVIEW

Chapter II will briefly discusses existing literature which related with this research title. This chapter has three sections. The first section describes the topic of data mining. The second section introduces the data clustering. The third section will explain the soft set theory.

2.1 Data Mining

The definition, examples and applications of Data Mining (DM) are presented in this section.

2.1.1 Definition of DM

Data mining (DM) is a data analyzing processes from multiple perspectives and angles, covert it into useful knowledge and information which can be used to benefit the system users and organization. Automated tools employing sophisticated algorithms used by data mining as purpose to find associations, hidden patterns, information repositories or data warehouse which in large amount of structures (Daniel.J.Power, Ramesh Sharda.

2007). Tasks of data mining are descriptive, like data describing by new pattern, and predict the model behavior based on data that available. Data mining consists of models fittings or patterns determination from data that observed. The models that been fitted act as inferred knowledge.

Generally, algorithm of data mining consist three elements below:

i. The model: model function (like clustering and classification) and its form representation (like neural networks). Parameters of model are determined from the data.

ii. The preference criterion: A fundamental for preference of one set or model parameters over another, based on data given. The criterion of the model to data is normally in form of goodness-of-fit function, and tempered by a smoothing expressions to prevent over fitting issues, or model generating with excess degrees of freedom to be restrict by the data.

iii. The search algorithm: the algorithm specification for finding specific parameters and models, given the model(s), preference criterion and data.

A particular algorithm of data mining normally is model/ search/ preference components instantiation. Usually functions of model in today data mining practice contain as show as below:

i. Classification: classifies item of data into one of few categorical classes that been predefined.

ii. Regression: scheme item of data to a real prediction valued variable.

iii. Clustering: scheme item of data into one of few clusters, where clusters are data items that natural groupings depend in probability density models or similarity metrics.

iv. Rule generation: extracts rules classifications from the data.

v. Discovering association rules: explain relationship that associate among attributes.

vi. Summarization: offers a strict description for a data subset.

vii. Dependency modeling: explains important dependencies among different variables.

viii. Sequence analysis: formulas sequential patterns. Example such as using analysis of time-series. The aim is to model the sequence and generating states process or to report and extract trends and deviation over time.

2.1.2 Examples of DM

In the investment and business field, plenty of companies use data mining as investment tool. Most of the companies do not explain their systems with except LBS Capital Management. The system is using genetic algorithm, neural networks and expert system to control portfolios which totaling around $600 million; Start from this system introduces in year 1993, the system has outperformed across the stock market.

Besides, another application of data mining example is DBMiner, a data mining system in data warehouses and relational databases. DBMiner is develops as the purpose for using in large data warehouses or relational databases that specific for multiple-level knowledge interactive mining. It carry out data mining in wide spectrum view, comprise of association, characterization, classification, comparison, clustering and prediction. DBMiner consist of several data mining techniques such as statistical analysis, attribute-oriented induction and OLAP (online analytical processing), mining which progressive deepening in multiple-level knowledge, and meta-rule guided base mining which provides an interactive, user-friendly environment of data mining with excellent performance. This system is testing on several platforms to large databases with outstanding performance, such as U.S. City-County Data Book, and Natural Science and Engineering Research Council of Canada (NSERC) research grant information system.

2.1.3 Applications of DM in Computer Science Fields

Knowledge discovery or data mining in databases touch upon extracting and clustering useful information from huge databases. This is happen because the need of recorded of increasing information in nowadays. Due to advance technologies today, most of the data are being stored in the computer databases form. Fitting to this issue, the huge amount of existence data is produce a need to transform the data into useful information. Many ways or methods are developed to address this issue and such ways or methods are been known as data mining. These methods allows the algorithms creation to be employ on knowledge of computer science to generate application that can handle process of data mining with the aim want to help in extracting the useful and important information from the large databases.

Data mining is used to solve and utilize different types of business problems, likes Enterprise Resource Planning (ERP), Supply Chain Management (SCM), and Customer Relationship Management (CRM). Large work amount are done in CRM to employ data mining with cooperate with Artificial Intelligence (AI) based customer segmentation and analysis. Symeonidis et al. (2003) synthesize techniques of data mining by clustering with agent technology to determine CRM and SCM modes and patterns, thus facilitate the implementation of ERP. Chen et al. (2005) using association rule mining for distribution centers to find out the batches order. Aitken et al. (2003) suggest a way that using a classification system to assort products and create appropriate and suitable supply chain outlines and strategies. Srinivasan et al. (1999) use a clustering algorithm to evaluate inventory decisions. Hong et al. (2005) employ clustering technique in grouping of customers which can then be used to support the selection of supplier.

Data mining been shown by Show-Jane Yen and Yue-Shi Lee in their paper is explain about the algorithm of incremental data mining to find and disclose the pattern of web access. These methods comprise the analysis and disclose the web useful information, and it is also been called as web mining. The problem that mentioned by them: 83% of the web systems only statically offer the homepages about the information services and page contents, with no count and consider the users behaviors. Due to this issue, the users waiting excess to find out the information needed by them. Therefore, by applying data mining, it is able to help in generating the useful and user-friendly traversal paths which suit the web user needs. This method indirectly faster the time needed for user to search and gather information.

2.2 Data Clustering

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

2.2.1 Definition

Cluster is determined as a data objects collections which are similar and close like to one another within the alike cluster but not alike to the other clusters objects. In another word, data clustering is the grouping of similar and alike data item into same clusters. The more detail data clustering definition is these clusters should reflect some mechanism at act in the domain from which data points and instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the reaming instances.

In order more comprehension about clustering, let explain the clustering concept in example format. For example, search the information on World Wide Web pages based on a given word: “eat”. This keyword will generate by the search engine and the results and outcome are shown. Results that consists the objects of the keyword “eat” will be show. For more sophisticated and high technology search engines, they will extend and enlarge the area of the search result by embody the keyword “eat” derivation like eating or “ate”. This is related because of the word “eating” concept pertained to same semantic group as “eat” and also, alike to “ate”, it is a form of grammatical transformation form for the word “eat”.

Both of the results are equally valid and relevant. Words like this type are called as homographs. Algorithms of clustering enable text documents which without labeling to be placed and located in each of these two sets. The aim is to determine the groupings results, which are cognate, close and similar.

2.2.2 Classification VS Clustering

Classification is nail down as the operation that wants to assign and distribute objects to predefined classes while clustering is determine as the task to imitate the data classification. This implies that in classification, predefined classification is not needed or required. Both of them are the pattern recognition central concept and significant knowledge discovery tools in learning process of modern machine. Classification consist of appointing data that input be into one or more pre-specified classes depend on attributes extraction of major features and the attributes analysis or processing of these attributes. Classification requires learning by example, or also named as supervised learning. Supervised learning is a procedure where the system aims to determine notion of classes' characterization that are jointly with examples of pre-classified. It presumes that the data are defined or labeled. Figure 2.2.1 illustrates the classification where observation or data (O) is mapped to the classes © with a descriptions or predefined patterns (P).

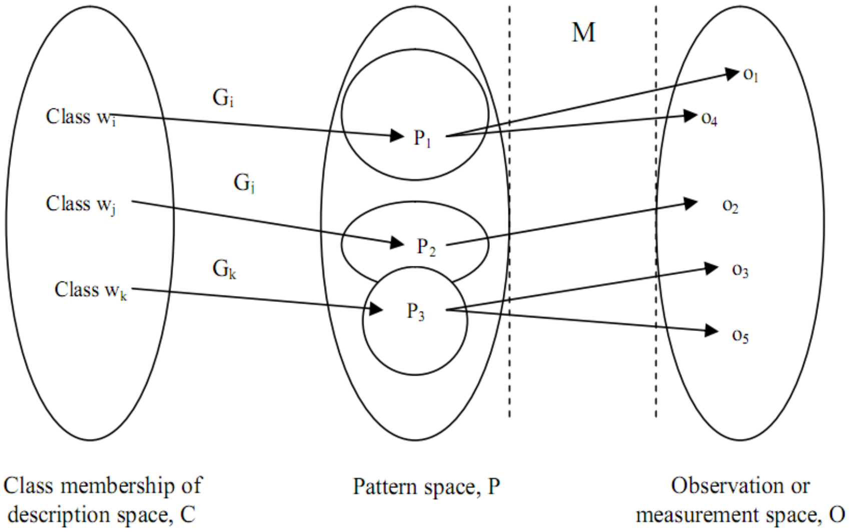

Figure 2.2.1: Classification

Nevertheless, in numerous cases, there are no obvious and distinct architecture in the data where the classes are not determined. There are no distinct number of classes and the connection between the characteristics or attributes of data are not obvious. In addition, the class-specific pattern attributes or characteristics when dealing with data over time, it can be modify continuously. So, clustering is used to settle and solve this issue. Clustering comprises group automatic identification for alike and similar specifications or object. The process is done by maximizing the similarity of inter-group and minimizing the similarity of intra-group (Marko Bohanec. 2009). The outcome show a number of clusters would form on the observation space or measurement. Then the data can be readily acquainted and be appointed to the appropriate and suitable clusters. This explanation is display in the following figure.

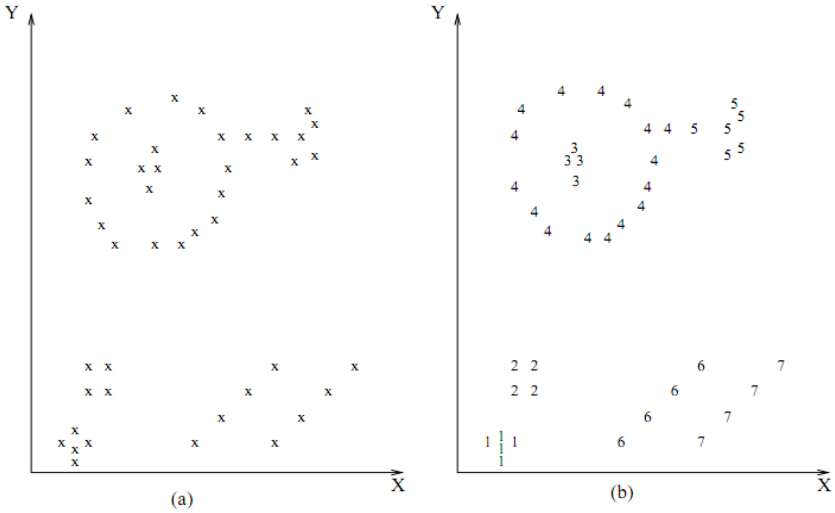

Figure 2.2.2: Clustering

Clustering demands unsupervised of data learning where the operation is only guided for data searching as interesting associations' purpose, and aim to group elements by postulating description and characterization of class for adequate numerous classes to cover and include all things in the data.

2.2.3 Clustering Techniques

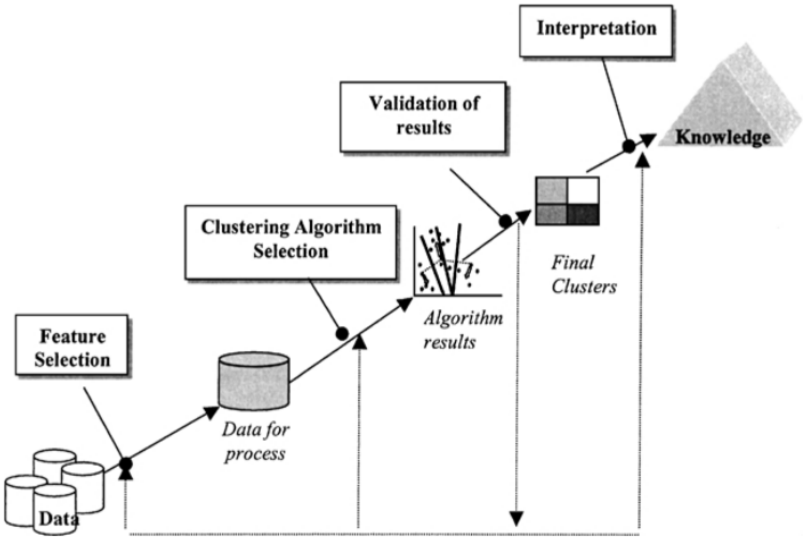

Figure 2.2.3: Clustering process

The clustering techniques comprise of the following procedures and process:

i. Feature selection: The process aim is to choose the proper features on which the clustering is to be carrying out so that as many as possible information can be acquired and obtained depend on task of the interest. So, the the data preprocessing may be required and necessary prior to the clustering task application, adoption and utilization.

ii. Clustering algorithm: This procedure consult as the algorithm choice which results and effect in the determination and definition of a well clustering scheme for a data set. A clustering criterion and a proximity measure mainly characterize a algorithm of clustering as well as its promptness and efficiency to determine a clustering scheme that suits the data set.

    - Proximity measure is a measure that quantifies how alike and similar of the two data points are. In other words, it imply how strong the relationship between the two data.

- Clustering criterion. The clustering criterion is to be determined in this step. Clusters types that are expected to happen in the dataset must be taken into account.

iii. Results validation: The clustering algorithm correctness results are verified using suitable techniques and criterions. Due to the clusters that been clustering are not predefined, the data final partition demands evaluation in most programs and applications.

iv. Results interpretation: The application area experts in most of the cases have to integrate and compare with the other experimental analysis and evidence clustering results with the aim to get the right conclusion.

2.2.4 Clustering on Numerical Dataset

Data can be partition into two types, numerical data and categorical data. Numerical data, or also named as quantitative data, is determined as data that consists of numerical counts or measures. The values of numerical data are used to calculate center measure, or to inspect the data spread. Clustering numerical data depend on a metric that decides the data pairs distance or the data pair's similarity level. The major metrics used are Canberra Metric, Euclidean distance, Mahalanobis distance and correlation coefficient. From these metrics mentioned above, two different techniques and approaches can be categorized: hierarchical and non- hierarchical clustering.

I. Hierarchical

The data are not divided into pre-defined clusters set but are connected to the closer group forming a single cluster including all objects (agglomerative methods) or partition in up to n number of clusters, each have a single object (divisive methods). The groups are connected in a dendrogram that shows the alike and similarity data structure and the groups' generated numbers rely on a cut parameter depend on the dendrogram generated user's analysis.

II. Non-hierarchical.

The most common algorithm of non-hierarchical is K-mean. It starts with an initial partition of the cases into k number of clusters and iterates with the aim to determine the best clusters, which is the one that minimizes the intra-clusters objects. Algorithm of K-means is very fast in drawback on depending on the pre-definition of the clusters number by the users. This is also considering as a problem because sometimes, the knowledge and information of the clusters ideal number is not accessible and available.

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

2.2.5 Clustering on Categorical Dataset

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

Previously, most of the algorithms clustering concentrate on numerical data. However, today the community of data mining is inundated with a large collection of categorical data such as those collected from health sectors, banks, biological data, web-log data, and market retailing data. Health sector and banking sector data are primarily mixed