Classification System for Incomplete Information Using Rough Set Approach

Ma Thinn Thinn Soe

Computer University (Mandalay)

Abstract

The incomplete information about attribute values may be the greatest obstacle to performing learning from examples. Rough set theory provides a powerful mathematical tool to deal with uncertainty and vagueness. Based on rough set theory, this system describes an approach for the classification and rule induction of incomplete information systems. In this system, attribute-value pair blocks are used as the main tool and these blocks are applied to construct characteristic sets, lower and upper approximations.

1.Introduction

Today, data mining is used in a vast array of areas, and numerous commercial data mining systems are available. Data mining also called Knowledge-Discovery Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc.

Classification is an important technique in data mining. Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. There are many classification methods such as decision tree induction, naïve bayesian classification, rule-based classification and rough set approach [3].

In most classification algorithms, it is assumed that information systems are complete. In real-life data, some attribute values are missing, imperfect or incomplete. Missing values in a data set can affect the performance of a classifier constructed using such a data set as a training sample [13].

Rough set theory provides powerful tools for classification from incomplete data. In this system, rough set theory is used to describe an approach for the classification and rule induction of incomplete information systems.

This thesis is organized with five chapters. Chapter 2 discusses classification methods and introduces rough set theory. Classification processes based on rough set theory are described in Chapter 3. Experimental results are presented in Chapter 4 and then it states system design and implementation. Finally Chapter 5 concludes and outlines directions for limitation and future extension.

2.Theoretical Background

Data mining concepts and techniques are hidden in large data set for uncovering interesting data patterns. The implementation methods discussed are particularly oriented towards the development of scalable and efficient data mining tools.

Data mining refers to extracting or “mining” knowledge from large amounts of data. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Many people treat data mining as synonym for another popularly used term, Knowledge Discovery in Database, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases [3].
Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: predictive and descriptive.

Predictive mining tasks perform inference on the current data in order to make predictions. The objective of these tasks is to predict the value of particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

Descriptive mining tasks characterize the general properties of the data in the database. The objective is to derive patterns that summarize the underlying relationship in data. Descriptive data mining tasks are often exploratory in nature and frequently require post processing techniques to validate and explain results. The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable [3].

Data mining functionalities are as follows:

1. Concept/Class Description
2. Mining Frequent Patterns, Associations, and Correlations
3. Classification and Prediction
4. Cluster Analysis
5. Outlier Analysis
6. Evolution Analysis

Back propagation learns by iteratively processing a set of training samples, comparing the network's prediction for each sample with the actual known class label. For each training sample, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual class. These modifications are made in the "backwards" direction, that is, from the output layer, through each hidden layer down to the first hidden layer. Although it is not guaranteed, in general the weights will eventually converge, and the learning process stops.

3. Rough Set Classification

Rough set theory, proposed in the 1980's by Pawlak, is widely applied in many areas such as machine learning, knowledge discovery in databases, expert systems, inductive reasoning, neural networks, decision systems, automatic classification, patterns recognition and learning algorithms [12].

Rough set theory is a special tool for handling the imprecise and incomplete data in information systems. In rough set theory, an information system can be presented in the form of a decision table, the rows and columns that correspond to cases (examples or objects) and characteristic conditions (a finite set of attributes and decision). The set of all cases and the set of all attributes is denoted by U and A respectively. B is a nonempty subset of the set A. An example of a decision table is shown in Table 3.1. The decision table may contain inconsistent examples, i.e. examples, characterized by the same values of attributes but the corresponding values of a decision are different. A decision table is incomplete when some attribute values are missing.

There are two main reasons for attribute values to be missing: either they are 'lost' or 'do not care' conditions. That is, originally the attribute value was known, however, due to a variety of reasons, currently

the value is not recorded and the decision to which concept a case belong was taken without that information.

In a decision table, there are three cases for missing values. These cases are as follows.

1. All missing attribute values are ‘lost values’.
2. All missing attribute values are ‘do not care’ conditions.
3. Some missing attribute values may be lost and some may be ‘do not care’ conditions [2].

A simple method for handling these cases is presented in this system. This system uses, as the main tool, attribute-value pair blocks. These blocks are used to construct characteristic sets, and lower and upper approximations for decision tables with missing attribute values. On the basis of the approximations, certain and possible rules are induced by using rule induction algorithm LEM2 (Learning from Examples Module, version2), a component of LERS (Learning from Examples based on Rough Set). Then, LERS classification method is used to classify new data.

4.Design and Implementation

The design of the system is presented in Figure 4.1. In this system, attribute-value pair blocks are used as the main tool. These blocks are used to construct characteristic sets, and lower and upper approximations for data set with missing attribute values. On the basis of the approximations, two corresponding sets of rules, certain and possible rules, are induced by using LEM2 algorithm, a component of LERS. Certain and possible rules may be further processed by using LERS classification algorithm. Finally, the accuracy of the classifier is measured.
This system is implemented by using C#. When the program starts, the main screen will appear as shown in Figure 4.2. This is the multi-documents interface application. The user is allowed to open to the corresponding forms from this main form. The main menu of the system consists of two categories. They are Data and Classification.

As rough set approach for missing case1 (Lost Value), there are five processes in rough set approach for missing case2 (Don’t Care Condition). Then, if the user clicks the Rough Set Approach for Missing Case2 (Don’t Care Condition) sub menu from Rough Set Approach menu, this system will perform these

processes. By using the presented rule induction algorithm, sixteen rules are generated and these rules are presented in Figure 4.14.

As rough set approach for missing case1 (Lost Value), there are five processes in rough set approach for missing case3 (Lost Value and Don't Care Condition). Then, if the user clicks the Rough Set Approach for Missing Case3 (Lost Value and Don't Care Condition) sub menu from Rough Set Approach menu, this system will perform these processes. The number of rules generated by the LEM2 (rule induction algorithm) are eighteen and these rules are presented in Figure 4.17.

5.Conclusions

Missing attribute values commonly exist in real world data set. Rough set approach provides powerful tools for classificatory analysis from imprecise and incomplete data and it does not require a preprocessing related to incomplete data (e.g. filling up incomplete data). Missing values in a data set can affect the performance of a classifier constructed, using such a data set as a training sample. There are three cases for missing values. This system uses rough set approach to handle these cases. In addition, the idea of an attribute-value pair block which is the main tool used in this system is both simple and useful. It is especially useful for incomplete decision tables. This system can be used to classify new mushroom whether it can be edible or poisonous by using LERS classification system. A conclusion is that interpreting missing attribute values as "***do not care***" conditions and "***lost*** and "***do not care***" conditions in a decision table" provides better results (better accuracy) than interpreting missing attribute values as "***lost***". This system is implemented by using C#.net on Microsoft.Net framework version 3.5 and SQL server 2005.

This system is implemented by using symbolic attributes and does not accept numerical attributes. When large data sets are used, this system takes longer processing time.

This system can be extended by using MLEM2 algorithm (Modified version of LEM2 algorithm) to induce rules from incomplete decision tables with numerical attributes. In this system, the mushroom data set with twenty-two attributes are used. Therefore, attribute reduction-an important issue in rough set theory can be used. The reduced set

of attributes provides the same quality of approximation as the original set of attributes. For computing the accuracy rate, k-fold cross validation and hold-out method can be used.

6.References

[1] Alaaeldin M.Hafez
"Knowledge Discovery in Databases"
Department of Computer Science and Automatic Control Faculty of Engineering, Alexandria University

[2] Grzymala_Busse, J.W. and Grzymala_Busse, Witold J.
"Chapter 1: Handling Missing Attribute Values"

[3] Jiawei Han and Micheline Kamber
"Data Mining: Concepts and Techniques"
Second Edition
Morgan Kaufmann Publishers, 2006

[4] Andrzej Skowron, Lech Polkowski and Pawlak, Z.
"Rough Sets: An Approach to Vagueness"

[5] Franco, L., Jerez, Jose M., Ignacio Molina, Subirats, Jose L., "Missing Data Imputation in Breast Cancer Prognosis", Proc of the 24th IASTED International Multi-Conference, Innsbruck, Austria, 2006

[6] Grzymala_Busse, J.W., "Data with Missing Attribute Values: Generalization of Indiscernibility Relation and Rule Induction", Springer_Verlag, Berlin, Heidelberg, 2004

[7] Grzymala_Busse, J.W., Goodwin, L.K., "A Closest Fit Approach to Missing Attribute Values in Preterm Birth Data", Springer_Verlag, Berlin, Heidelberg, 1999

[8] Grzymala_Busse, J.W. and Hu, M.
"A Comparison of Several Approaches to Missing Attribute Values in Data Mining", 2001

[9] Grzymala_Busse, J.W. and Siddhaye, S., "Rough Set Approaches to Rule Induction from Incomplete Data", the 10th International Conference on IPMU, Perugia, Italy, 2004

[10] Grzymala_Busse, J.W. and Wang, Chien Pei B., "Classification

Methods in Rule Induction", Proc of the Fifth Intelligent Information Systems Workshop, Deblin, Poland, June 2–5, 1996

[11] Jiye Li and Nick Cercone, "Assigning Missing Attribute Values Based on Rough Sets Theory", Fellow, IEEE

[12] Pawlak, Z., "Rough sets", International Journal of Computing and Information Science, Vol. 11, pp.341-356, 1982

[13] Thangavel, K., Pethalakshmi, A. and Jaganathan, P., "A Novel Reduct Algorithm for Dimensionality Reduction with Missing Values Based on Rough Set Theory", Medwell Onlilne, 2006

[14] Zhu, X. and Wu, X., "Cost-Constrained Data Acquisition for intelligent Data Preparation", IEEE Transactions on Knowledge and Data Engineering, 2005

[15] http:// www.ics.uci.edu/~mlearn/MLRepository.html.