UNIVERSITI PUTRA MALAYSIA

NEW LEARNING MODELS FOR GENERATING CLASSIFICATION RULES BASED ON ROUGH SET APPROACH

LUAI ABDEL LATEEF AL SHALABI

FSKTM 2000 2

NEW LEARNING MODELS FOR GENERATING CLASSIFICATION RULES BASED ON ROUGH SET APPROACH

By

LUAI ABDEL LATEEF AL SHALABİ

Thesis Submitted in Fulfilment of the Requirement for the Degree of Doctor of Philosophy in the Faculty of Computer Science and Information Technology Universiti Putra Malaysia

September 2000

Dedicated to my father; Abdel Lateef,

my mother; Faidah,

my wife; Samah and the family

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements of the degree of Doctor of Philosophy

NEW LEARNING MODELS FOR GENERATING CLASSIFICATION RULES BASED ON ROUGH SET APPROACH

By

LUAI ABDEL LATEEF AL SHALABI

September 2000

Chairman: Ramlan Mahmod, Ph.D.

Faculty: Computer Science and Information Technology

Data sets, static or dynamic, are very important and useful for presenting real life features in different aspects of industry, medicine, economy, and others. Recently, different models were used to generate knowledge from vague and uncertain data sets such as induction decision tree, neural network, fuzzy logic, genetic algorithm, rough set theory, and others. All of these models take long time to learn for a huge and dynamic data set. Thus, the challenge is how to develop an efficient model that can decrease the learning time without affecting the quality of the generated classification rules. Huge information systems or data sets usually have some missing values due to unavailable data that affect the quality of the generated classification rules. Missing values lead to the difficulty of extracting useful information from that data set. Another challenge is how to solve the problem of missing data.

Rough set theory is a new mathematical tool to deal with vagueness and uncertainty. It is a useful approach for uncovering classificatory knowledge and building a classification rules. So, the application of the theory as part of the learning models was proposed in this thesis.

Two different models for learning in data sets were proposed based on two different reduction algorithms. The split-condition-merge-reduct algorithm (SCMR) was performed on three different modules: partitioning the data set vertically into subsets, applying rough set concepts of reduction to each subset, and merging the reducts of all subsets to form the best reduct. The enhanced-split-condition-merge-reduct algorithm (ESCMR) was performed on the above three modules followed by another module that applies the rough set reduction concept again to the reduct generated by SCMR in order to generate the best reduct, which plays the same role as if all attributes in this subset existed. Classification rules were generated based on the best reduct.

For the problem of missing data, a new approach was proposed based on data partitioning and function mode. In this new approach, the data set was partitioned horizontally into different subsets. All objects in each subset of data were described by only one classification value. The mode function was applied to each subset of data that has missing values in order to find the most frequently occurring value in each attribute. Missing values in that attribute were replaced by the mode value.

The proposed approach for missing values produced better results compared to other approaches. Also, the proposed models for learning in data sets generated the

classification rules faster than other methods. The accuracy of the classification rules by the proposed models was high compared to other models.

Abstrak tesis yang dikemukan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MODEL PEMBELAJARAN BARU UNTUK MENJANA PETUAPENGKELASAN BERDASARKAN PENDEKATAN SET KASAR

Daripada

LUAI ABDEL LATEEF AL AHALABI

September 2000

Pengerusi: Ramlan Mahmod, Ph.D

Fakulti: Sains Komputer dan Teknologi Maklumat

Set-set data, statik atau dinamik, adalah sangat penting dan berguna untuk memaparkan fitur realiti hidup dalam berlainan aspek di bidang industri, perubatan, ekonomi dan lain-lain. Kebelakangan ini, pelbagai model telah digunakan untuk menjana pengetahuan daripada set-set data yang kabur dan tidak pasti seperti pokok keputusan induksi, rangkaian neural, logik kabur, algoritma genetik, teori set kasar dan sebagainya. Kesemua model tersebut mengambil masa yang panjang untuk belajar bagi set-set data yang besar dan dinamik. Maka, cabarannya ialah bagaimana untuk membangunkan satu model yang berkesan yang boleh mengurangkan masa pembelajarannya tetapi tidak menjejaskan kualiti petua klasifikasi yang terjana. Maklumat yang besar atau set-set data yang besar selalu mempunyai nilai yang tercicir akibat daripada ketidaksempurnaan data yang menyebabkan kerumitan untuk mengekstrak maklumat yang berfaedah daripada set-set data itu.

Teori set kasar adalah alat matematik baru untuk mengatasi masalah kekaburan dan ketidakpastian. Ia adalah pendekatan yang berguna untuk mendapatkan pengetahuan pengkelasan dan membina satu petua pengkelasan. Maka, aplikasi teori set kasar sebagai sebahagian daripada model pembelajaran telah diperkenalkan dalam tesis ini.

Dua model pembelajaran berlainan dalam set-set data telah dicadangkan berdasarkan kepada dua algoritma reduksi berlainan. Algoritma SCMR telah dilakukan ke atas tiga modul berlainan: pembahagian set-set data kepada subset, aplikasikan konsep reduksi set kasar kepada setiap subset, dan gabungkan hasil reduksi semua subset untuk membentuk reduksi terbaik. Model ESCMR telah digunakan ke atas tiga modul di atas dan diikuti dengan satu modul lain yang mengaplikasikan sekali lagi konsep reduksi set kasar ke atas reduksi yang terhasil daripada SCMR, untuk menjana reduksi terbaik. Reduksi ini memainkan peranan yang sama seperti semua atribut di dalam subset. Petua pengkelasan dijana berdasarkan reduksi terbaik.

Pendekatan baru telah dicadangkan untuk mengatasi masalah ini. Dalam pendekatan baru ini, set-set data telah dikategorikan secara mendatar kepada subset yang berlainan. Semua objek dalam setiap subset data ditakrifkan oleh hanya satu nilai klasifikasi. Fungsi mod telah diaplikasikan ke atas setiap subset data yang mengandungi nilai tercicir untuk mencari nilai yang mempunyai frekuensi tertinggi dalam setiap atribut. Nilai tercicir dalam atribut itu digantikan dengan nilai mod

Pendekatan yang dicadangkan untuk nilai tercicir ini menghasilkan keputusan yang lebih baik berbanding dengan pendekatan lain. Model yang dicadangkan dalam pembelajaran set-set data, telah dapat menjana peraturan klasifikasi lebih cepat daripada kaedah lain. Ketepatan petua pengkelasan oleh model yang dicadangkan adalah lebih tinggi berbanding dengan model-model yang lain.

ACKNOWLEDGEMENTS

Praise to Allah for giving me the strength and patience to complete this moderate work.

Firstly, I would like to thank my supervisor Dr. Ramlan Mahmod, deputy dean, Faculty of Computer Science and Information Technology, for his supervision and invaluable discussion. My thanks also to my co-supervisors, Dr. Abdul Azim Abd ghani, dean, Faculty of Computer Science and Information Technology and Dr. Md. Yazid Mohd. Saman for their invaluable discussion and help.

I am certainly grateful to Prof. Ohmann, Theoretical Surgery Unit, Department of General and Trauma Surgery, Heinrich-Heine-University, Dusseldrof, Germany, for providing the original Acute Abdominal Pain data set.

I also wish to express my thanks to the Faculty of Computer Science and Information Technology, Post Graduate Office, and Library, University Putra Malaysia, for providing assistance and good environment while I was pursuing my research work.

Special thanks are due to my friends and colleagues who have made some assistance in the development of this research. From list I would like to mention my friends Ayman Oklat and Ziad Abu Kaddorah.

Finally, I would like to thank my parents, wife, sisters, and brother for their patience, understanding and moral support.

Luai Abdel Lateef Al Shalabi

September, 2000

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BR	Best Reduct
ESCMR	Enhanced Split Condition Merge Reducts
FOL	First Order Logic
ID3	Induction of Decision Trees
IS	Information System
ITRule	Information Theoretic Rule Induction
PRISM	Programming In Statistical Modeling
RoughDAS	Rough sets based Data Analysis System
RoughIAN	Rough Information Analysis
RSDA	Rough Set Data Analysis
SCMR	Split Condition Merge Reduct
TDIDT	Top-Down Induction of Decision Tree

CHAPTER 1

INTRODUCTION

1.1 Background

Data mining or knowledge discovery from databases is a form of machine discovery where the discovered knowledge is represented in a high level language. It is capable of discovering domain knowledge from given examples. However, the theory of knowledge discovery is still under development and few existing methods are practical. The type of rule or pattern that exists in data depends on the domain. Discovery systems have been applied to real databases in medicine, astronomy, the stock market, and many others.

Data mining has come to refer to the process of analysing data and generating new knowledge that is previously hidden and unseen. The overall goal is to create a simplified model of the domain under study. Various techniques for data mining have been employed, mostly from the area of inductive learning, with different forms of knowledge representation such as weights in artificial neural network or nodes in a decision tree.

A well-known and widely employed inductive learning algorithm is ID3 (Quinlan, 1979; 1986). ID3 and subsequent versions (Quinlan, 1993) based their decisions on statistical measures of the information entropy and knowledge is represented as a decision tree which may easily be converted to a set of rules. An alternative method is based on rough sets which is based on the theory of sets and

topology. Rough set theory was introduced by Pawlak (Pawlak, 1982) and since then a number of applications have been reported in diverse fields such as medical diagnosis, conflict analysis, and process control (Slowinski, 1992). Rough set can be used for the purpose of generating *If... Then* rules and/or as a technique for eliminating redundant information prior to the use of, say, artificial neural networks.

Rough set theory, introduced by Pawlak in 1982 (Pawlak, 1991; 1995), is a new mathematical tool dealing with vagueness and uncertainty. It has proved its soundness and usefulness in many real life applications. Rough set theory offers effective methods that are applicable in many branches of AI. The idea of rough set consists on approximation of a set by a pair of sets called lower and upper approximations of the set. The definition of the approximation follows from an indiscrenibility relation between elements of the sets, called objects. Objects are described by attributes of a qualitative nature.

Rough set can be a useful tool for pre-processing data for generating classification rules. Applying its concepts to an application at hand reduces the number of attributes and the complexity of the classification rules (Al-Shalabi et al., 1999b).

1.2 Objectives of the Study

The main goal of this dissertation is to propose new models based on time and cost for learning classification rules from data sets. Solving the problem of repeating a learning process to the whole original data set if it changes is a special case of the proposed models. This learning process takes some time to generate new

classification rules. As we know most data sets are frequently changed. This kind of data sets is called "dynamic" data sets. Time consuming is the important disadvantages of the existing systems that are used to generate classification rules especially if the data set is huge. In this study, medical diagnosis has been applied.

The objectives of the study can be derived from the main goal, and they include:

1. To produce a new approach for pre-processing input with missing data to the proposed models.

2. To produce new algorithms for reducing a number of attributes of a data set based on the rough set theory and the partitioning of the data set.

3. To produce new models for the discovery of accurate classification rules from data sets based on the new algorithms.

1.3 Significance of the Study

Data sets, static or dynamic, are very important and useful for presenting real life features in different aspects of industry, medicine, economy, and others. They act as store of information that can answer most questions. In order to gain the most knowledge from these data sets, a special technique to sieve and clean these data is used and then correct knowledge is generated. Building an expert system is one of these techniques. It is considered one of the most important and early used systems

that can generate knowledge from databases. Recently, many other techniques have been used to generate knowledge from databases or data sets such as ID3, neural network, fuzzy logic, genetic algorithm, rough set theory, and others. We may say that all these methods have the capability of generating knowledge requested from databases or data sets. But these methods seem to have learning time and learning complexity problems especially if the data set is dynamic. So, this raises a question of how to develop an efficient tool that can help the task and decrease the learning time without affecting the quality of the rules generated.

This study proposes new different models to approach this problem. These have the idea of partitioning the data set vertically into two or more subsets. Then a rough set theory is applied to each subset in the new space in order to find the best reduct for each of them. The best reduct contains less number of attributes, strong discernibly attributes, low cost, and some other features. The combination of all reducts is done in the higher level and as a result the knowledge presented by classification rules are discovered.

The concept of rough sets has been proposed as a new mathematical tool to deal with uncertain and imprecise data, and it seems to be of significant importance to AI and cognitive sciences (Slowiniski, 1992). Using this tool to approach the problem of data reduction and data dependency has emerged as a powerful technique in the application of expert systems, decision support systems, machine learning, and pattern recognition.