UNIVERSITI PUTRA MALAYSIA

TWOFOLD INTEGER PROGRAMMING MODEL FOR IMPROVING ROUGH SET CLASSIFICATION ACCURACY IN DATA MINING

WALID SAEED.

FSKTM 2005 3

TWOFOLD INTEGER PROGRAMMING MODEL FOR IMPROVING ROUGH SET CLASSIFICATION ACCURACY IN DATA MINING

By

WALID SAEED

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

September 2005

اعوذ بالله من الشيطان الرجيم

(قال رب اشرح لي صدري* ويسر لي امري* واحلل عقدة من لساني* يفقهو قولي)

سورة الاسراء- أية (25-28)

This thesis is dedicated to my parents, my wife
and to anyone believes that we have to do strong effort for our nation.

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

TWOFOLD INTEGER PROGRAMMING MODEL FOR IMPROVING ROUGH SET CLASSIFICATION ACCURACY IN DATA MINING

By

WALID SAEED

September 2005

Chairman: Associate Professor Hj. Md. Nasir Sulaiman, Ph.D.

Faculty: Computer Science and Information Technology

The fast growing size of databases has resulted in a great demand for tools capable of analyzing data with the aim of discovering new knowledge and patterns. These tools will hopefully close the gap between the steady growth of information and the escalating demand to understand and discover the value of such knowledge. These tools are known as Data Mining (DM).

One aims of DM is to discover decision rules for extracting meaningful knowledge. These rules consist of conditions over attribute value pairs called the descriptions, and decision attributes. Therefore generating a good decision model or classification model is a major component in many data mining researches. The classification approach basically produces a function that maps data item into one of several predefined classes, by way of inputting training dataset and building a model of the class attribute based on the rest of the attributes.

This research undertakes three main tasks. The first task is to introduce a new rough model for minimum reduct selection and default rules generation, which is known as a Twofold Integer Programming (TIP). The second task is to enhance rules accuracy based on the first task, while the third task is to classify new objects or cases.

The TIP model is based on translation of the discernibility relation of a Decision System (DS) into an Integer Programming (IP) model, resolved by using the branch and bound search method in order to generate the full reduct of the DS. The TIP model is then applied to the reduct to generate the default rules, which in turn are used to classify unseen objects with a satisfying accuracy.

Apart from introducing the TIP model, this research also addressed the issues of missing values, discretization and extracting minimum rules. The treatment of missing values and discretization are being carried out during the preprocessing stage. The extraction of minimum rules operation is conducted after the default rules have been generated in order to obtain the most useful discovered rules.

Eight datasets from machine learning repositories and domain theories are tested by the TIP model. Total rules number, rules length and rules accuracy for the generation rules are recorded. The accuracy for rules and classification resulted from the TIP method are compared with other methods such as Standard Integer Programming (SIP) and Decision Related Integer Programming (DRIP) from Rough Set, Genetic Algorithm (GA), Johnson reducer, Holte1R method, Multiple Regression (MR), Neural Network (NN), Induction of Decision Tree Algorithm

(ID3) and Base Learning Algorithm (C4.5); all other classifiers that are mostly used in the classification tasks.

Based on the experiment results, the classification method using the TIP approach has successfully performed rules generation and classification tasks as required during a classification operation. The outcome of a considerably good accuracy is mainly due to the right selection of relevant attributes. This research has proven that the TIP method has shown the ability to cater for different kinds of datasets and obtained a good rough classification model with promising results as compared with other commonly used classifiers.

This research opens a wide range of future work to be considered, which includes applying the proposed method in other areas such as web mining, text mining or multimedia mining; and extending the proposed approach to work in parallel computing in data mining.

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

MODEL PENGATURCARAAN DIGIT DWIARAS BAGI MEMPERBAIKI KETEPATAN KLASIFIKASI SET KASAR DALAM PERLOMBONGAN DATA

Oleh

WALID SAEED

September 2005

Pengerusi **: Profesor Madya Hj. Md. Nasir b. Hj. Sulaiman, Ph.D.**

Fakulti **: Sains Komputer dan Teknologi Maklumat**

Kadar peningkatan saiz pangkalan data yang semakin meninggi telah mewujudkan keperluan ke atas alat menganalisis data yang bertujuan untuk menemui corak dan pengetahuan baru. Alatan-alatan ini diharapkan dapat menutup jurang antara peningkatan pertumbuhan maklumat dan keperluan bagi memahami dan mencari pengetahuan baru yang amat berharga.

Salah satu tujuan perlombongan data adalah untuk menemui petua-petua keputusan bagi mengekstrak pengetahuan baru yang bermakna. Petua-petua ini mengandungi syarat-syarat ke atas pasangan nilai atribut yang dikenali sebagai deskripsi dan atribut kata putus. Oleh yang demikian, pembinaan model keputusan atau klasifikasi merupakan komponen terpenting dalam kebanyakan penyelidikan perlombongan data. Teknik-teknik pengklasifikasian pada dasarnya menghasilkan satu fungsi yang memetakan item data kepada beberapa jenis kelas yang telah didefinasikan terlebih dahulu, dengan cara memasukkan set data latihan dan membina model bagi kelas atribut berdasarkan atribut-atribut yang lain.

Proses penyelidikan ini dibahagikan kepada tiga tugasan utama. Tugasan pertama adalah untuk memperkenalkan satu model kasar baru untuk pilihan pengurang minimum dan penjanaan petua-petua lalai yang dikenali sebagai Pengaturcaraan Digit Dwi-Aras (TIP). Tugasan kedua adalah untuk meningkatkan ketepatan petua-petua hasil daripada tugasan pertama, sementara tugasan ketiga adalah untuk mengklasifikasikan objek-objek atau kes-kes baru.

Model TIP adalah berasaskan kepada penterjemahan bagi hubungan nyata untuk satu sistem keputusan (DS) kepada satu model Pengaturcaraan Digit (IP) yang diselesaikan menggunakan kaedah model gelidahan iaitu cabang dan pantul dengan tujuan menjana pengurang lengkap bagi DS tersebut. TIP kemudiannya diaplikasikan ke atas pengurang terbaik bagi menjana petua lalai yang akhirnya digunakan untuk mengklasifikasikan objek-objek terselindung dengan ketepatan yang memuaskan.

Selain memperkenalkan model TIP, penyelidikan ini turut mengutarakan isu-isu nilai yang hilang, diskretizasi dan pengekstrakan petua minimum. Baik pulih nilai-nilai hilang dan diskretizasi dijalankan sewaktu paras pra-pemprosesan. Operasi pengekstrakan petua minimum dijalankan selepas pengurangan petua lalai telah dijana bagi mendapatkan petua yang paling berguna atau menarik.

Lapan set data yang diperoleh daripada simpanan pembelajaran mesin dan teori domain diuji oleh model TIP ini. Jumlah bilangan panjang dan ketepatan petua semasa penjanaan peraturan direkodkan. Ketepatan peraturan dan klasifikasi terhasil daripada kaedah TIP dibandingkan dengan kaedah-kaedah lain seperti *Standard Integer Programming* (SIP) dan *Decision Related Integer Programming* (DRIP) daripada *Rough Set*, *Genetic Algorithm* (GA), *Johnson*, kaedah *Holte1R*, *Multiple Regression* (MR), *Neural Network* (NN), *Induction of Decision Tree Algorithm* (ID3) dan *Base Learning Algorithm* (C4.5); kesemuanya merupakan pengklasifikasi utama dalam tugasan pengklasifikasian.

Berdasarkan keputusan eksperimen, kaedah pengklasifikasian TIP telah berjaya melaksanakan penjanaan petua-petua dan kerja-kerja pengklasifikasian yang diperlukan semasa operasi klasifikasi. Keputusan ketepatan yang menggalakkan adalah hasil daripada pemilihan yang betul terhadap atribut yang relevan. Penyelidikan ini telah membuktikan bahawa kaedah TIP berupaya menangani set-set data yang berlainan dan telah berjaya memperoleh model klasifikasi kasar yang baik berserta keputusan yang memberangsangkan setanding dengan pengklasifikasi yang lain.

Penyelidikan ini membuka jalan kepada pelbagai isu untuk kajian masa hadapan, termasuk penggunaan kaedah yang dicadangkan dalam bidang-bidang lain seperti perlombongan web, teks atau multimedia; dan penambahan kepada kaedah cadangan ini agar boleh digunakan untuk perlombongan data dalam pengkomputeran selari.

ACKNOWLEDGEMENTS

In the name of *Allah*, He is all Merciful, Most Gracious and Most Compassionate and who is the Creator of all knowledge for eternity. We beg for peace and blessings upon our Master the beloved Prophet Muhammad (Peace and Blessings be Upon Him) and his progeny, companions and followers. All grace and thanks belong to Almighty *Allah*.

I wish to extend my deepest appreciation and gratitude to the supervisory committee led by *Assoc. Prof. Dr. Hj. Md. Nasir Bin Sulaiman* and committee members, *Prof. Hj. Hassan Selamat, Assoc. Prof. Dr. Mohd Othman* and *Dr. Azuraliza Abu Bakar* for their virtuous guidance, sharing of intellectual experiences and in giving me the vital to undertake the numerous aspects of this study.

Special appreciation to my parents for their loves and prayers and my wife for making the best of my situation. My thanks are also extended to my friends and colleagues, sharing experiences throughout the years.

WALID SAEED

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF TABLES cont.

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANNs	Artificial Neural Networks
ATPG	Automatic Test Pattern Generation
AUS	Australian Credit Card
BCO	Breast Cancer Dataset
C4.5	Base Learning Algorithm
CA	Classification Accuracy
CLEV	Cleveland Heart Disease
CNF	Conjunctive Normal Form
CQA	Congressional Quarterly Almanac
DBMS	Database Management Systems
DM	Data Mining
DRIP	Decision Related Integer Programming
DS	Decision Systems
EM	Exhaustive Method
EMR	Extracting Minimum Rules
GA	Genetic Algorithms
GERM	German Credit
H1R	Holte1R Reducer
ID3	Induction of Decision Tree
IP	Integer Programming
IRIS	Iris Plants Dataset
IS	Information System
IT	Information Technology

LIST OF ABBREVIATIONS cont.

GR	Johnson Reducer
KDD	Knowledge Discovery in Database
LYM	Lymphography Dataset
ML	Machine Learning
MLP	Multilayer Perceptron
MR	Multiple Regression
NN	Neural Networks
NT	New Thyroid Disease
RS	Rough Set
RVD	Real Value Discretization
SAT	Propositional SATisfiability
SIP	Standard Integer Programming
TIP	Twofold Integer Programming
UCI	University of California, Irvine
VDM	Value Difference Metrics
VOTING	Voting Records Database

CHAPTER 1

INTRODUCTION

1.1 Background

With the growing amount of information in the world, knowledge discovery and data mining in large databases become the most interesting topic for researchers and many major companies in the Information Technology (IT) area. Aside from the steady growth of information, there is also a mounting demand for tools that are capable of analyzing patterns from large amounts of data in search of invaluable knowledge and hidden information.

Knowledge Discovery in Databases (KDD) is getting to be very important and has grown recently. The huge amounts of data collected and stored might contain some information, which could be useful, but it is not easy to recognize, nor trivial to obtain it. There is no human being capable of sifting through such amounts of data and even some existing algorithms are inefficient when trying to solve this problem. The process of knowledge discovery is generally defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data (Mollestad, 1997). Thuraisingham (1999) defined Data Mining (DM) as the process of posing various queries and extracting useful information, patterns and trends often previously unknown from large quantities of data possibly stored in databases.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information, which can be used to increase revenue, cut costs, or both. It has gained considerable attention among practitioners and researchers as evidenced by the number of publications, conferences and application reports. The growing volume of data that is available in a digital form has accelerated this interest, and classification is one of the most common data mining tasks. Data mining relates to other areas, including machine learning, cluster analysis, regression analysis and neural networks (Kusiak, 2001).

A classification process produces a function that maps a data item onto one of several predefined classes, by means of inputting training data set and building the model for a class attribute based on the rest of the attributes. Learning accurate classifiers from pre-classified data is still a very active research in the field of machine learning and data mining. Data mining researchers use classifiers to identify and classify important objects within a data repository. Classification is particularly useful when a dataset contains examples that can be used as the basis for future decision-making.

Although the classification is an important and useful process in knowledge representation systems, the processing time increases rapidly as the size of the knowledge base increases (Kim, 1993; Bazan *et al.*, 2002). Han and Kamber (2001) defined classification as a process of finding a set of models or functions.

Classification is one of the most common data mining tasks, seems to be a human imperative (Berry and Linoff, 2004). In this work, the rough classification model introduced is structured based on the rough analysis method to extract the important rules, which are used during the classification operation. This new approach aids in reducing the dataset and can be used as the source for future decision making to arrive at a high quality of knowledge.

The integer programming problem is a linear integer programming problem where all variables are restricted to take values of either 0 or 1. This problem is considered unlikely that there exists an efficient algorithm for solving it (Hillier and Lieberman, 1989). Most IPs are solved by using the approach of branch-and-bound. Branch-and-bound methods find the optimal solution to an IP by efficiently enumerating the points in a subproblems feasible region (Winston, 1994).

1.2 Problem Statement

Our knowledge is incomplete and problems are waiting to be solved. We can address the holes in our knowledge and those unresolved problems by asking relevant questions and then seeking answers through systematic research (Leedy and Ormrod, 2001).

KDD is an active research area with the promise of a high payoff in many business and scientific applications. The grand challenge of knowledge discovery in database is to automatically process large amounts of raw data, identify the most significant meaningful patterns, and present this knowledge in an appropriate for achieving the user's goal (Hu, 1995).