



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 به اخرج لي حذري * ويسر لي امري * واجلل عتقة من لساني * بفتقمو قولي)

سورة الاسراء - آية (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 tinggi 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 didefinisikan 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 tugas utama. Tugas 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). Tugas kedua adalah untuk meningkatkan ketepatan petua-petua hasil daripada tugas pertama, sementara tugas 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, diskretisasi dan pengekstrakan petua minimum. Baik pulih nilai-nilai hilang dan diskretisasi 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 *Holte IR*, *Multiple Regression* (MR), *Neural Network* (NN), *Induction of Decision Tree Algorithm* (ID3) dan *Base Learning Algorithm* (C4.5); kesemuanya merupakan pengklasifikasi utama dalam tugas 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



I certify that an Examination Committee met on 1st September 2005 to conduct the final examination of Waled Saeed Abdo on his Doctor of Philosophy thesis entitled “Twofold Integer Programming Model for Improving Rough Set Classification Accuracy in Data Mining” in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

RAMLAN MAHMUD, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

ABDUL AZIM ABD GHANI, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

HAMIDAH IBRAHIM, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

SAFAAI DERIS, PhD

Professor
Faculty of Computer Science and Information System
Universiti Teknologi Malaysia
(External Examiner)



GULAM RUSUL RAHMAT ALI, PhD
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: **22 NOV 2005**



This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirements for the degree of Doctor of Philosophy. The members of the Supervisory Committee are as follows:

Md. Nasir Bin Sulaiman, PhD

Associate Professor

Faculty of Computer Science and Information Technology

Universiti Putra Malaysia

(Chairman)

Mohd. Hasan Selamat, M.Phi.

Associate Professor

Faculty of Computer Science and Information Technology

Universiti Putra Malaysia

(Member)

Mohd Othman, PhD

Associate Professor

Faculty of Computer Science and Information Technology

Universiti Putra Malaysia

(Member)

Azuraliza Abu Bakar, PhD Lecturer

Faculty of Information Science and Technology

Universiti Kebangsaan Malaysia

(Member)



AINI IDERIS, PhD

Professor/Dean

School of Graduate Studies

Universiti Putra Malaysia

Date: 8 DEC 2005



DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.



WALID SAEED

Date: 8 / 10 / 2005

TABLE OF CONTENTS

DEDICATION	ii
ABSTRACT	iii
ABSTRAK	vi
ACKNOWLEDGEMENTS	x
APPROVAL	xi
DECLARATION	xiii
LIST OF TABLES	xvii
LIST OF FIGURES	xix
LIST OF ABBREVIATIONS	xx

CHAPTER

1 INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Objectives of the Research	5
1.4 Scope of the Research	5
1.5 Contributions of the Research	7
1.6 Overview of Thesis	7
2 LITERATURE REVIEW	10
2.1 Introduction	10
2.2 Data Mining (DM)	11
2.2.1 Data Mining Steps	11
2.2.2 Data Mining Functions	13
2.2.3 Classification Modeling	16
2.2.4 Data Mining Challenges	18
2.2.5 Performance Criteria	21
2.2.6 Propositional SATisfiability (SAT) Problem	22
2.3 Rough Approach	24
2.3.1 Information System (IS)	25
2.3.2 Decision System (DS)	26
2.3.3 Set Approximation	27
2.3.4 Indiscernibility Relation	29
2.3.5 Equivalence Classes	29
2.3.6 Discernibility Matrix	31
2.3.7 Reduct	32
2.3.8 Generation of Rules	34
2.4 SIP/DRIP Approach	35
2.4.1 SIP/DRIP Outlines	35
2.4.2 Standard Integer Programming (SIP) Algorithm	37
2.4.3 Decision Related Integer Programming (DRIP) Algorithm	38
2.4.4 Critical Analysis	39
2.4.5 SIP/DRIP Algorithms Compared to Previews Methods	40



2.5	Integer Programming (IP) Model	41
2.5.1	Integer Programming (IP)	41
2.5.2	Branch-and-Bound	42
2.5.3	A Branch-and-Bound Algorithm for Binary Linear Programming	43
2.6	Summary	45
3	RESEARCH METHODOLOGIES	47
3.1	Introduction	47
3.2	Preparation of Datasets	47
3.2.1	Acquisition of Datasets	48
3.2.2	Selected Datasets	49
3.2.3	Missing Values Treatment	53
3.2.4	Generation of Training and Test Dataset	53
3.2.5	Data Discretization	54
3.2.6	Equivalence Classes Generation	55
3.2.7	Data Preprocessing Phases	55
3.3	System Design and Implementation	56
3.3.1	Overall System Architecture	58
3.3.2	Design of Experiments	59
3.3.3	Reduct Generation	62
3.3.4	Rules Generation	62
3.3.5	Rules Accuracy Computation	63
3.3.6	Classification Operation	63
3.4	Statistical Measurements	64
3.5	Comparison Methods Design	66
3.5.1	Rules Results Comparison	66
3.5.2	Classification Comparison	67
3.6	Summary	67
4	TWOFOLD INTEGER PROGRAMMING (TIP) MODEL	69
4.1	Introduction	69
4.2	Definitions and Related Terms	70
4.3	Data Preprocessing	71
4.3.1	Handling Missing Values	72
4.3.2	Real Value Discretization (RVD) Algorithm	74
4.4	Proposed TIP Model	78
4.4.1	TIP Model	79
4.4.2	Clause Database	84
4.4.3	Complexity of the TIP Model	86
4.5	Generation of Default Rules	87
4.6	Extracting the Minimum Rules (EMR) Algorithm	89
4.7	TIP Model in the Framework	92
4.8	Summary	94
5	EXPERIMENTAL RESULTS	95
5.1	Introduction	95
5.2	Rules Accuracy Experiments	96
5.2.1	Wisconsin Breast-Cancer (BCO)	97
5.2.2	New-Thyroid Disease Dataset (NT)	98
5.2.3	Experimental Results Discussion	99



5.3	Comparison of TIP Approach Against Other Classifiers	101
5.3.1	Australian Credit Card (AUS)	102
5.3.2	Cleveland Heart Disease (CLEV)	102
5.3.3	Lymphography (LYM) Dataset	103
5.3.4	German Credit (GERM) Dataset	104
5.3.5	Experimental Results Discussion	105
5.4	Experiments of Objects Classification Accuracy	108
5.4.1	Experimental Results of Objects Classification Accuracy	109
5.4.2	Comparison of the TIP Against Other Two Methods	110
5.4.3	Comparative Results and Discussion	112
5.5	Summary	113
6	CONCLUSIONS AND FUTURE WORK	115
6.1	Introduction	115
6.2	Conclusions	116
6.3	Capabilities of the Proposed Model	117
6.4	Future Research	118
	BIBLIOGRAPHY	120
	APPENDICES	130
	PUBLICATIONS	175
	BIODATA OF THE AUTHOR	176



LIST OF TABLES

Table		Page
2.1	An Information System for the Study Categories of Students	26
2.2	A Decision System for the Income Categories of Students	27
2.3	Equivalence Classes of the Decision System	30
2.4	Numerical Representation of Equivalence Classes	31
2.5	The Discernibility Matrix of the DS	32
2.6	Decision System	39
2.7	IP Model Using SIP Algorithm	40
2.8	The Results of SIP/DRIP Compared with Other Methods	41
3.1	Generation of Training and Test Dataset	54
4.1	VDM and RVD Discretization	77
4.2	Decision System	82
4.3	The Decision System After Dropping Attribute a	83
4.4	The Decision System After Dropping Attribute b	83
4.5	The Clause Database for the Reduct Generation	84
4.6	IP Model for Reduct Generation	85
4.7	The Clause Database for Rules Generation	88
4.8	TIP Reducts for Rules Generation	88
4.9	Rules Generation Using TIP	89
5.1	Characteristics of Selected Datasets	97
5.2	TIP Rules Accuracy for BCO Dataset	98
5.3	TIP Rules Accuracy for NT Dataset	99
5.4	Rules Accuracy for Algorithmic Methods	99



LIST OF TABLES cont.

Table		Page
5.5	Characteristics of Selected Datasets	101
5.6	Rules Accuracy for AUS Dataset	102
5.7	Rules Accuracy for CLEV Dataset	103
5.8	Rules Accuracy for LYM Dataset	104
5.9	Rules Accuracy for GERM Dataset	105
5.10	Summary of Rules Accuracy for Extracted Rules	108
5.11	Objects Classification Accuracy	109
5.12	The Characteristics of IRIS and VOTING Datasets	110
5.13	Objects Classification Accuracy Results of TIP, ID3 and C4.5	111



LIST OF FIGURES

Figure		Page
2.1	Set Approximation	28
2.2	SIP Algorithm	37
2.3	DRIP Algorithm	38
3.1	Data Preprocessing Phases	57
3.2	The Architecture of the Proposed System	59
3.3	Experiments Steps	61
4.1	Real Value Discretization (RVD) Algorithm	78
4.2	The TIP Model to Compute the IP for Full Reduct	81
4.3	The TIP Model to Compute the IP for Default Rules Generating	82
4.4	Extracting the Minimum Rules/Reduct Algorithm	91
4.5	TIP Steps Framework	93
5.1	Rules Accuracy of Algorithmic Methods	100
5.2	Rules Accuracy of Selected Methods	108
5.3	Objects Classification Accuracy for LYM and CLEV Datasets	110
5.4	Difference Objects Classification Accuracy	111



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).

