UNIVERSITI PUTRA MALAYSIA

PROPOSITIONAL SATISFIABILITY METHOD IN ROUGH CLASSIFICATION MODELING FOR DATA MINING

AZURALIZA ABU BAKAR

PROPOSITIONAL SATISFIABILITY METHOD IN ROUGH CLASSIFICATION MODELING FOR DATA MINING

By

AZURALIZA ABU BAKAR

Thesis Submitted in Fulfillment of the Requirements for the degree of
Doctor of Philosophy in the Graduate School
Universiti Putra Malaysia

January 2002

I asked for Strength… and Allah gave me difficulties to make me strong.
I asked for Wisdom… and Allah gave me problems to solve.
I asked for Prosperity… and Allah gave me brain and brawn to work.
I asked for Courage… and Allah gave me danger to overcome.
I asked for Favours… and Allah gave me opportunities.
I received nothing I wanted… I received everything I needed.
My prayer has been answered.

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

PROPOSITIONAL SATISFIABILITY METHOD IN ROUGH CLASSIFICATION MODELING FOR DATA MINING

By

AZURALIZA ABU BAKAR

January 2002

Chairman : Md. Nasir Sulaiman, Ph.D.

Faculty : Computer Science and Information Technology

The fundamental problem in data mining is whether the whole information available is always necessary to represent the information system (IS). The goal of data mining is to find rules that model the world sufficiently well. These rules consist of conditions over attributes value pairs called description and classification of decision attribute. However, the set of all decision rules generated from all conditional attributes can be too large and can contain many chaotic rules that are not appropriate for unseen object classification. Therefore the search for the best rules must be performed because it is not possible to determine the quality of all rules generated from the information systems. In rough set approach to data mining, the set of interesting rules are determined using a notion of

reduct. Rules were generated from reducts through binding the condition attribute values of the object class from which the reduct is originated to the corresponding attribute. It is important for the reducts to be minimum in size. The minimal reducts will decrease the size of the conditional attributes used to generate rules. Smaller size of rules are expected to classify new cases more properly because of the larger support in data and in some sense the most stable and frequently appearing reducts gives the best decision rules.

The main work of the thesis is the generation of classification model that contains smaller number of rules, shorter length and good accuracy. The propositional satisfiability method in rough classification model is proposed in this thesis. Two models, **S**tandard **I**nteger **P**rogramming (*SIP*) and **D**ecision **R**elated **I**nteger **P**rogramming (*DRIP*) to represent the minimal reduct computation problem were proposed. The models involved a theoretical formalism of the discernibility relation of a decision system (DS) into an Integer Programming (IP) model. The proposed models were embedded within the default rules generation framework and a new rough classification method was obtained. An improved branch and bound strategy is proposed to solve the SIP and DRIP models that pruned certain amount of search. The proposed strategy used the conflict analysis procedure to remove the unnecessary attribute assignments and determined the branch level for the search to backtrack in a non-chronological manner.

Five data sets from UCI machine learning repositories and domain theories were experimented. Total number rules generated for the best classification model is recorded

where the 30% of data were used for training and 70% were kept as test data. The classification accuracy, the number of rules and the maximum length of rules obtained from the SIP/DRIP method was compared with other rough set method such as Genetic Algorithm (GA), Johnson, Holte1R, Dynamic and Exhaustive method. Four of the datasets were then chosen for further experiment. The improved search strategy implemented the non-chronological backtracking search that potentially prunes the large portion of search space. The experimental results showed that the proposed SIP/DRIP method is a successful method in rough classification modeling. The outstanding feature of this method is the reduced number of rules in all classification models. SIP/DRIP generated shorter rules among other methods in most dataset. The proposed search strategy indicated that the best performance can be achieved at the lower level or shorter path of the tree search. SIP/DRIP method had also shown promising across other commonly used classifiers such as neural network and statistical method. This model is expected to be able to represent the knowledge of the system efficiently.

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

KAEDAH KEPUASAN USULAN DALAM PEMODELAN PENGELASAN KASAR UNTUK PERLOMBONGAN DATA

Oleh

AZURALIZA ABU BAKAR

Januari 2002

Pengerusi : Md. Nasir Sulaiman, Ph.D.

Fakulti : Sains Komputer dan Teknologi Maklumat

Masalah utama dalam melombongi data ialah sama ada keseluruhan maklumat yang ada sentiasa perlu untuk mewakili satu sistem maklumat. Matlamat melombongi data ialah mencari petua yang memodelkan alam dengan baik. Petua-petua ini mengandungi syarat-syarat ke atas pasangan nilai atribut yang dipanggil deskripsi dan pengelasan atribut kataputus. Walau bagaimanapun, set semua petua yang dijanakan dari semua atribut bersyarat boleh menjadi terlalu besar dan boleh mengandungi banyak petua yang berserabut yang tidak dipaerlukan untuk pengelasan objek baru. Oleh itu pencarian petua yang terbaik mesti dilaksanakan kerana adalah tidak mungkin untuk menentukan kualiti semua petua yang dijana daripada sistem maklumat. Dalam pendekatan set kasar ke atas perlombongan data, set petua yang menarik ditentukan menggunakan notasi reduksi. Petua dijana daripada reduksi dengan mengikat nilai atribut syarat kepada satu kelas objek daripada reduksi yang diperolehi, keatas atribut berkaitan. Adalah penting untuk reduksi bersaiz minimum. Reduksi minima mengurangkan saiz atribut syarat yang digunakan untuk menjana petua. Saiz petua yang lebih kecil dijangka dapat mengelaskan kes-kes baru dengan lebih baik kerana sokongan yang lebih besar ke atas data.

Kerja utama tesis ini ialah penjanaan model pengelasan yang mengandungi bilangan petua yang sedikit, petua yang lebih pendek dan ketepatan yang baik. Kaedah kepuasan usulan dalam pemodelan pengelasan kasar dicadangkan di dalam tesis ini. Dua model, Standard Integer Programming (SIP) dan Decision Related Integer Programming (DRIP) untuk mengira reduksi minimal dibincangkan. Model-model tersebut melibatkan satu formalisma teoritikal keatas hubungan ketaksamaan bagi satu sistem kataputus kepada satu model pengaturcaraan integer. Model yang dicadangkan kemudiannya di bina di dalam rangka kerja penjanaan petua lalai menghasilkam satu kaedah pengelasan kasar yang baru. Satu strategi cabang dan sempadan dicadangkan untuk menyelesaikan model-model SIP/DRIP untuk mengurangkan sebilangan jumlah carian. Strategi yang dicadangkan menggunakan tatacara analisis knoflik untuk memangkas umpukan atribut yang tidak perlu dan menentukan paras cabang untuk carian jejak kebelakang dengan keadaan tidak bertertib.

Lima set data dari *UCI machine learning repositories and domain theories* diuji. Jumlah petua yang dijana untuk model pengelasan terbaik direkod iaitu 30% daripada data digunakan untuk latihan dan 70% lagi disimpan sebagai data ujian. Ketepatan pengelasan, bilangan petua dan panjang maksimum petua yang diperolehi daripada kaedah SIP/DRIP dibandingkan dengan kaedah set kasar yang lain seperti kaedah algoritma genetik, Johnson, Holte1R, Dynamic dan Exhaustive. Empat daripada set data kemudian dipilih untuk eksperimen seterusnya. Set data latihan kemudiannya dieksperimen ke atas strategi carian yang dicadangkan. Strategi pembaikan carian tersebut melaksanakan carian jejak kebelakang tak tertib yang berpotensi mengurangkan sebahagian besar ruang carian sambil mengekalkan carian lengkap. Hasil eksperimen menunjukkan kaedah SIP/DRIP merupakan kaedah yang berjaya dalam memodelkan pengelasan kasar. Satu fitur yang terbaik ialah bilangan petua yang minimum dalam semua model pengelasannya. Kaedah SIP/DRIP juga menjana petua yang terpendek dalam kebanyakan set data. Strategi carian yang dicadangkan menunjukkan pencapaian yang terbaik boleh diperolehi pada paras carian pohon yang lebih bawah atau laluan yang lebih pendek. Kaedah SIP/DRIP juga menunjukkan hasil yang setanding dan lebih baik berbanding dengan kaedah pengelasan yang biasa digunakan iaitu rangkaian neural dan kaedah statistik. Model ini dijangka dapat mewakili pengetahuan sistem secara cekap.

ACKNOWLEDGEMENTS

In the name of *Allah*, the most merciful and most compassionate. Praise to *Allah s.w.t.* for granted me strength, courage, patience and inspirations in completing this work.

My deepest appreciation and gratitude to the supervisory committee leads by *Assoc. Prof. Dr. Md. Nasir Sulaiman* and committee members, *Assoc. Prof. Dr. Mohamed Othman* and *Assoc. Prof. Mohd Hasan Selamat* for their virtuous guidance, sharing their intellectual experiences and giving their motivation and support that lead the way in so many aspect of the research work.

I would like to thank the members of Knowledge System Group IDI, Norweigian University of Science and Technology and Dr. J.P.M Silva from Cadence European Laboratories/INESC for providing materials, papers and book chapters. I acknowledge the influences of the *Soft Computing and Mathematical Technology Research Team* for their stimulating discussions and ideas. For the financial support, I am grateful to *Universiti Utara Malaysia* for the scholarship, study leave and allowances.

Special appreciation to my parents for their loves and prayers and my family for making the best of my situation. Sincere thanks to friends and colleagues, sharing experiences throughout the years.

Azuraliza Abu Bakar

January 2002

TABLE OF CONTENTS

DEDICATION ii
ABSTRACT iii
ABSTRAK vi
ACKNOWLEDGEMENT ix
APPROVAL SHEETS x
DECLARATION FORM xii
LIST OF TABLES xvii
LIST OF FIGURES xx
LIST OF ABBREVIATIONS xxi

CHAPTER

1 INTRODUCTION

1.1 Background 1
1.2 Problem Statement 4
1.3 Objective of the Research 5
1.4 Scope of the Research 6
1.5 Research Methodology 7
1.6 Contributions of the Research 8
1.7 Organization of the Thesis 9

2 LITERATURE REVIEW

2.1 Data Mining 11
2.2 Data Mining Problems and Challenge 13
2.3 Performance Criteria 16
2.4 Classification Modeling in Data Mining 17
 2.4.1 Statistical Classifier 18
 2.4.2 Neural Classifier 20
 2.4.3 Rough Classifier 22
2.5 Rulebase Classification Methods 24
2.6 Compact Classification Models 27
2.7 Rough Classification Modeling 29
 2.7.1 Propositional Satisfiability Problem (SAT) in Minimal Reduct Computation 30
 2.7.2 SAT Based Branch and Bound Method 32
2.8 Summary 34

3 ROUGH CLASSIFICATION MODELING

3.1 Fundamentals of Rough Sets 35
3.2 Knowledge Representation 36
3.2.1 Information Systems (IS) 36
3.2.2 Decision System (DS) 38
3.3 Discerning Objects 39
3.3.1 Indiscernibility Relation 40
3.3.2 Equivalence Class 41
3.3.3 Discernibility Matrix 43
3.3.4 Discernibility Matrix Modulo D 44
3.3.5 Discernibility Function 45
3.4 Reduct 46
3.4.1 Reduct and Prime Implicants 46
3.4.2 Reduct Computation Methods 48
3.5 Generation of Rules 50
3.6 Default Rules Generation Framework 53
3.6.1 Degree of Membership 55
3.6.2 Projections of Decision System 56
3.6.3 Search in the Lattices 57
3.6.4 Pruning the Search in the Lattice 61
3.6.5 Rules Measurements 64
3.7 Summary 67

4 METHODOLOGY OF THE PROPOSED METHOD

4.1 Introduction 68
4.2 Proposed SIP/DRIP Models 69
4.2.1 Definitions 70
4.2.2 *SIP* Model 73
4.2.3 *DRIP* Model 75
4.3 Branch and Bound Search Method in Solving *SIP* and *DRIP* 78
4.3.1 Feasible Solution 79
4.3.2 Upper and Lower Bound Calculation 79
4.3.3 Conflict Analysis 80
4.3.4 Lower Bound Conflict 81
4.3.5 Logical Conflict 81
4.3.6 Upper Bound Conflict 82
4.3.7 Standard Branch and Bound Search Strategy 82

4.4 The Improved Branch and Bound Search Method 86
4.4.1 Handling the Lower Bound Conflicts 87
4.4.2 Handling the Logical Conflicts 89
4.4.3 The Improved Search Algorithm 91
4.4.4 Complexity of the Improved Search Algorithm 93

4.5 SIP and DRIP Models in Rough Classification Modeling 93
4.5.1 The Clause Database 96
4.5.2 Minimal Literal Length 97
4.5.3 Clause Subsumption 98
4.5.4 Pruning the Search in the Lattices 100
4.5.5 The Algorithm of SIP/DRIP in the Framework 103
4.6 Summary 105

5 EXPERIMENTS AND OBSERVATION

5.1 Introduction 106
5.2 Experiment Remarks of the Proposed Method 106
5.3 The Case Studies 108
5.3.1 Australian Credit Card Approval (*AUS*) 111
5.3.2 Cleveland Heart Disease(*CLEV*) 115
5.3.3 Lymphography(*LYM*) 120
5.3.4 Breast Cancer (*BCO*) 123
5.3.5 German Credit (*GERM*) 127
5.4 Experimental Results on the Improved Branch and Bound Search Strategy 132
5.5 Comparison with Neural and Statistical Classifiers 136
5.5.1 Experimental Results on Neural Classifier 137
5.5.2 Comparative Results 140
5.6 Summary 144

6 CONCLUSIONS AND FUTURE WORK

6.1 Concluding Remark 146
6.2 Capabilities of the Proposed Method 148
6.3 Future Research 150

BIBLIOGRAPHY 152

APPENDICES

A-1 Reducts and Prime Implicant of the IS/DS 162
A-2 Reduct Computation Methods 165
B-1 AUS Data Description 170
B-2 CLEV Data Description 173
B-3 LYM Data Description 177
B-4 BCO Data Description 181
B-5 GERM Data Description 183
C Data Discretization with Boolean Reasoning 187

D Classification Rules : General Information 194

D-1 AUS Rules 195

D-2 CLEV Rules 201

D-3 LYM Rules 206

D-4 BCO Rules 210

D-5 GERM Rules 212

E Survey of Classification Accuracy Results on Other Machine Learning Algorithms 215

F Summary of Results of SIP/DRIP with Other Methods for all datasets 220

VITA 221

LIST OF TABLES

Table		Page
3.1	An Information System of the Income Categories of Workers	37
3.2	A Decision System of the Income Categories of Workers	39
3.3	Equivalence Classes Obtained from a Decision System	42
3.4	Numerical Representation of the Decision	42
3.5	The Discernibility Matrix	43
3.6	The Discernibility Matrix Modulo D	44
3.7	Reduct of a Decision System	48
3.8	Rules Generation	52
3.9	The Projection of Classes	58
3.10	Definite and Default Rules	61
4.1	CNF Boolean Formulation of the Discernibility Relation	73
4.2	CNF Boolean Formulation of the *Decision Related* Discernibility Relation	76
4.3	SIP/DRIP Reducts	84
4.4	Rules generation using SIP/DRIP reducts	84
4.5	The Clause Database	97

Table **Page**

5.1 SIP/DRIP Classification Accuracy for AUS dataset 112

5.2(a) Comparison of SIP/DRIP to Other Methods (AUS) 112

5.2(b) Comparison of SIP/DRIP to Other Methods (AUS) 113

5.3 Summary of Experimental Results (AUS) 114

5.4 AUS Best Model Rules Description 115

5.5 The Classification Accuracy for CLEV dataset 116

5.6(a) Comparison of SIP/DRIP to Other Methods (CLEV) 117

5.6(b) Comparison of SIP/DRIP to Other Methods (CLEV) 117

5.7 Summary of Experimental Results (CLEV) 118

5.8 CLEV Best Model Rules Description 119

5.9 SIP/DRIP Classification Accuracy for LYM Dataset 120

5.10(a) Comparison of SIP/DRIP to Other Methods (LYM) 121

5.10(b) Comparison of SIP/DRIP to Other Methods (LYM) 121

5.11 Summary of Experimental Results (LYM) 122

5.12 LYM Best Model Rules Description 123

5.13 The Classification Accuracy for BCO dataset 124

5.14(a) Comparison of SIP/DRIP to Other Methods (BCO) 124

5.14(b) Comparison of SIP/DRIP to Other Methods (BCO) 125

5.15 Summary of Experimental Results(BCO) 126

5.16 BCO Best Model Rules Description 126

5.17 SIP/DRIP Classification Accuracy for GERM Dataset 128

5.18(a) Comparison of SIP/DRIP to Other Methods (GERM) 129

5.18(b) Comparison of SIP/DRIP to Other Methods (GERM) 129

5.18(c) Comparison of SIP/DRIP to Other Methods (GERM) 130

5.18(d) Comparison of SIP/DRIP to Other Methods (GERM) 130

5.19 Summary of Experimental Results (GERM) 131

5.20 GERM Best Model Rules Description 132

5.21 Improved Search Strategy (NCHR) vs Chronological Search Strategy (CHR) for AUS Dataset 133

5.22 Improved Search Strategy (NCHR) vs Chronological Search Strategy (CHR) for CLEV Dataset 134

5.23 Improved Search Strategy (NCHR) vs Chronological Search Strategy (CHR) for LYM Dataset 134

5.24 Improved Search Strategy (NCHR) vs Chronological Search Strategy (CHR) for BCO Dataset 135

5.25 NN Classification Accuracy for AUS Dataset 138

5.26 NN Classification Accuracy for CLEV Dataset 139

5.27 NN Classification Accuracy for LYM Dataset 139

5.28 Classification Accuracy for SIP/DRIP, NN, MR (AUS) 141

5.29 Classification Accuracy for SIP/DRIP, NN, MR (CLEV) 141

5.30 Classification Accuracy for SIP/DRIP, NN, MR (LYM) 142

5.31 AVG and MAX Classification Accuracy 143

LIST OF FIGURES

Figure		Page
3.1	Rules Generation Using Reducts	51
3.2	Default Rules Generation Framework	54
3.3	The search in the Lattices	59
4.1	The Branch and Bound Search Strategy	83
4.2	The Improved Branch and Bound Search Strategy	91
4.3	SIP/DRIP in Default Rules Generation Framework	94
5.1	The SIP/DRIP in Rough Classification Modeling Process	110

LIST OF ABBREVIATIONS

AUS	Australian Credit Card
BCO	Breast Cancer of Ontology Institute
CLEV	Cleveland Heart Disease
CNF	Conjunctive Normal Form
DRIP	Decision Related Integer Programming
DS	Decision System
GA	Genetic Algorithms
GERM	German Credit
IP	Integer Programming
IS	Information System
LYM	Lymphography
MLP	Multi Layer Perceptron
MR	Multiple Regression
NN	Neural Network
SAT	Satisfiability
SIP	Standard Integer Programming

CHAPTER 1

INTRODUCTION

1.1 Background

As the amount of information in the world is steady increasing, there is a growing demand for tools for analyzing the information with the aim of finding patterns in terms of implicit dependencies in data. Realizing that much of the collected data will not be handled or even seen by human beings, systems that are able to generate summaries from large amount of information will be important currently. Although several statistical techniques for data analysis were developed long ago, advanced techniques for intelligent data analysis are not yet mature. As a result, there is a growing gap between data generating and data understanding. On a high level, different formalisms for machine learning that uses the notion of knowledge are expected to be able to extract interesting and useful patterns from large collection of data. Several different paradigms have been developed over the years.

The concept of *knowledge discovery* has recently been brought to attention of the business community. One main reason for this is the need of the general recognition that perform the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. The term *data mining* is used to denominate the process of automatic extraction of information in knowledge discovery process. In this thesis, the extracted knowledge is represented as a set of propositional rules that that can be formally defined as a relationship between a set of attribute values and a decision. Rules are the easiest pattern for human interpretation and understandings. A set of propositional rules forms a model, which explains how different values for the attribute lead to different decisions. Two main tasks for which the model is useful are *prediction* and *classification*. Basically prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Classification is the process of finding the common properties among different objects and classifying the objects into classes.

Classification is probably the most well known data mining problem. It has been widely studied by researchers in the artificial intelligence (AI) field. The database community focuses on searching efficient classification algorithms. Their work involves either developing new and efficient classification algorithms (Agrawal et al., 1992;1993) or further advancing the existing AI techniques for example extracting rules in "if … then …" form that can be applied to large database. In the real world applications, the number of attributes of a dataset could be very large. It is common for a class label of an object depends only on the values of a few attributes. In such cases, presenting the original data with all the attributes into a classifier may confuse the classifier to generate unnecessary complex classification rules. The knowledge extraction in the classification context is the process of selecting the most important attributes from the information systems or a dataset.

There are few advantages if a small set of attributes can be determined before the actual objects are fed into a classifier. First, the time required to extract rules is reduced because a smaller input data set compared to the original one. For example, if neural networks are used for classification, less number of attributes means less number of nodes at the input layer and less training time required. For decision tree based classifiers, less number of attributes means less computation required for classification criteria and less comparison is made. Secondly, as the attributes that do not contribute to the classification are removed, the rules generated by the classifiers are expected more concise than the original dataset used (Lu et al., 1995).

With a number of the classification algorithms available, a natural task is developed appropriate performance metrics that can be used to compare the goodness of these algorithms. Classification accuracy is the most important performance metric of a classification method. It is defined as the ratio between the number of correctly classified objects and the total number of objects in the test set. However when two different methods are applied to the same classification problem, they most likely will extract different set of rules. Certain criteria need to be introduced to compare the goodness of the extracted rules. An ideal rule set has a minimum number of rules and each rule is as short as possible. In practice, it is quite often that a rule set contains fewer rules but they usually have more conditions. Therefore, shorter rules should rather be generated although they will not be perfect on the known cases there are good chances of good classification quality when classifying new cases.