UNIVERSITI PUTRA MALAYSIA

CASE SLICING TECHNIQUE FOR FEATURE SELECTION

OMAR A.A. SHIBA.

FSKTM 2004 6

CASE SLICING TECHNIQUE FOR FEATURE SELECTION

By

OMAR A. A. SHIBA

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

June 2004

بسم الله الرحمن الرحيم

ومن يتق الله يجعل له مخرجا * ويرزوقه من حيث لايحتسب ومن يتوكل على الله فهو حسبه
إن الله بالغ أمره قد جعل الله لكل شيء قدرا

سورة الطلاق أية (2,3)

This thesis is dedicated to my parents and my family.

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

CASE SLICING TECHNIQUE FOR FEATURE SELECTION

By

OMAR A. A. SHIBA

June 2004

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

Faculty: **Computer Science and Information Technology**

One of the problems addressed by machine learning is data classification. Finding a good classification algorithm is an important component of many data mining projects. Since the 1960s, many algorithms for data classification have been proposed. Data mining researchers often use classifiers to identify important classes of objects within a data repository.

This research undertakes two main tasks. The first task is to introduce slicing technique for feature subset selection. The second task is to enhance classification accuracy based on the first task, so that it can be used to classify objects or cases based on selected relevant features only. This new approach called Case Slicing Technique (CST). Applying to this technique on classification task can result in further enhancing case

classification accuracy. Case Slicing Technique (CST) helps in identifying the subset of features used in computing the similarity measures needed by classification algorithms.

CST was tested on nine datasets from UCI machine learning repositories and domain theories. The maximum and minimum accuracy obtained is 99% and 96% respectively, based on the evaluation approach. The most commonly used evaluation technique is called *k*-cross validation technique. This technique with $k = 10$ has been used in this thesis to evaluate the proposed approach.

CST was compared to other selected classification methods based on feature subset selection such as Induction of Decision Tree Algorithm (ID3), Base Learning Algorithm K-Nearest Nighbour Algorithm (k_NN) and Naïve Bayes Algorithm (NB). All these approaches are implemented with RELIEF feature selection approach.

The classification accuracy obtained from the CST method is compared to other selected classification methods such as Value Difference Metric (VDM), Pre–Category Feature Importance (PCF), Cross-Category Feature Importance (CCF), Instance-Based Algorithm (IB4), Decision Tree Algorithms such as Induction of Decision Tree Algorithm (ID3) and Base Learning Algorithm (C4.5), Rough Set methods such as Standard Integer Programming (SIP) and Decision Related Integer Programming (DRIP) and Neural Network methods such as the Multilayer method.

Based on the results of the experiments, the best performance could be achieved using the slicing technique. It also gave promising results across other commonly used classifiers such as machine learning, neural network and statistical methods. Likewise, the technique is able to enhance the classification accuracy.

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

TEKNIK HIRISAN KES BAGI PEMILIHAN FITUR

Oleh

OMAR A. A. SHIBA

Jun 2004

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

Fakulti : **Sains Komputer dan Teknologi Maklumat**

Salah satu masalah yang dibincangkan oleh pembelajaran mesin adalah pengklasifikasian data. Mencari satu algoritma klasifikasi yang baik adalah komponen penting dalam banyak projek perlombongan data. Sejak tahun 1960-an, pelbagai algoritma pengklasifikasian data telah dikemukakan. Penyelidik perlombongan data biasanya menggunakan pengkelas untuk mengenal pasti kelas penting bagi objek dalam satu simpanan data.

Kajian ini mengandungi dua misi utama. Misi pertama adalah untuk memperkenalkan teknik penghirisan untuk pemilihan subset fitur. Misi kedua adalah untuk menambahkan kejituan pengklasifikasian berdasarkan tugas yang pertama, dengan itu ianya dapat digunakan untuk mengkelaskan objek-objek atau kes-kes berdasarkan fitur berkaitan sahaja. Pendekatan baharu ini dikenali sebagai Teknik Penghirisan Kes (CST). Pembaharuan dalam teknik ini boleh menyumbang dalam kejituan pengklasifikasian kes.

CST berupaya mengenal pasti subset kepada fitur yang digunakan untuk membilang ukuran kesamaan yang diperlukan oleh algoritma klasifikasi.

CST telah diuji ke atas tujuh set data daripada teori domain dan simpanan pembelajaran mesin UCI. Kejituan maksimum dan minimum yang diperoleh berdasarkan pendekatan penilaian ini adalah masing-masing 99% dan 96%. Teknik penilaian yang paling biasa digunakan adalah teknik pengesahsahihan silang-k. Teknik ini dengan k =10 telah digunakan dalam tesis ini untuk menilai pendekatan yang telah dicadangkan.

CST juga dibandingkan dengan kaedah-kaedah klasifikasi yang lain berdasarkan subset pemilihan fitur seperti induksi Algoritma Pokok keputusan, Algoritma Pembelajaran Dasar, Algoritma k-jiran terdekat dan Algoritma Naïve Bayes. Kesemua kaedah ini telah dimplementasikan dengan pendekatan pemilihan fitur RELIEF.

Kejituan pengklasifikasian yang diperoleh daripada kaedah CST dibandingkan dengan kaedah pengklasifikasian yang lain seperti Metrik Perbezaan Nilai (*Value Difference Metric*-VDM), Kepentingan Fitur Pra-Kategori (*Pre-Category Feature Importance-PCF*), Kepentingan Fitur Kategori-Silang (*Cross-Category Feature Importance-CCF*), Algoritma Berdasarkan-Contoh (*Instance-Based Algorithm-IB4*), Algoritma Pepohon Keputusan (*Decision Tree Algorithms*) seperti Aruhan Algoritma Pepohon Keputusan (*Induction of Decision Tree Algorithm-ID3*) dan Algoritma Pembelajaran Asas (*Base Learning Algorithm-C4.5*), kaedah Set Kasar (*Rough Set Methods*) seperti Pengaturcaraan Integer Piawai (*Standard Integer Programming*-SIP) dan Pengaturcaraan Integer Berkaitan Keputusan (*Decision Related Integer Programming-*

DRIP) serta kaedah Rangkaian Neural (*Neural Network*) seperti kaedah Berbilang Lapisan (*Multilayer*).

Berdasarkan keputusan eksperimen, prestasi terbaik boleh diperoleh menggunakan teknik hirisan ini. Ia juga memberikan keputusan yang lebih baik berbanding pengkelas lain yang biasa digunakan seperti pembelajaran mesin, rangkaian neural dan kaedah statistik. Teknik ini mampu meningkatkan kejituan klasifikasian.

ACKNOWLEDGEMENTS

All praise is due to Almighty Allah as He is all Merciful, Most Gracious and Most Compassionate and it is He who has gathered all knowledge in His Essence 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 Sulaiman* and committee members, *Assoc. Prof. Dr. Hj. Ali Mamat* and *Assoc. Prof. Dr. Hjh. Fatimah Dato' Ahmad* for their virtuous guidance, sharing of intellectual experiences and in giving me the vital to undertake the numerous aspects of this study.

I acknowledge with deep gratitude my parents for their love and prayers. They have always encouraged me and guided me to seek knowledge and never limited my aspirations at any time.

My wife deserves my unending gratitude for her assistance in looking after our children and making our home quiet and a suitable place for study.

My thanks are also extended to colleagues for sharing experiences over the years.

My heartfelt thanks to all.

Omar A. A. Shiba

January 2004

TABLE OF CONTENTS

3 CLASSIFICATION TASK

4 FEATURES SELECTION METHODS

5 THE PROPOSED METHOD

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ASCII	American Standard Code for Information Interchange
AUS	Australian Credit Card
BCO	Breast Cancer of Ontology Institute
C4.5	Base Learning Algorithm
CB	Case Base
CBR	Case-Based Reasoning
CCF	Cross-Category Features
CLEV	Cleveland Heart Disease
CNF	Conjunctive Normal Form
CRX	Credit Card Application
CST	Case Slicing Technique
DBMS	Database Management Systems
DNA	Primate Slice-Junction Gene Sequences
DRIP	Decision Related Integer Programming
DS	Decision Systems
GA	Genetic Algorithms
GERM	German Credit
HEPA	Hepatitis
IB4	Instance Based Algorithm
ID3	Induction of Decision Tree
IP	Integer Programming
IRIS	Iris Plants Dataset
IS	Information System
K-CV	K-Cross Validation
KDD	Knowledge Discovery in Database
K-NN	K-Nearest Neighbour
LYMP	Lymphography
MAP	Maximum A Priori
ML	Machine Learning
MLP	Multilayer Perceptron
MR	Multiple Regression
NB	Naïve Bayes
NN	Neural Networks
PCF	Pre-Category Features
RS	Rough Set
SDG	System Dependence Graph
SIP	Standard Integer Programming
SQL	Structured Query Language
VDM	Value Difference Metric
VOTING	Voting Records Database

CHAPTER 1

INTRODUCTION

1.1 Background

Advances in database technologies and data collection techniques including barcode reading, remote sensing, satellite telemetry, etc. have accumulated huge amounts of data in large databases. This explosive growth in data creates the necessity of knowledge/information discovery from data which has led to the promising emergence of a new field called data mining or knowledge discovery in databases (KDD) (Fayyad *et al.*, 1996(a); 1996(b); Holsheimer and Siebes, 1994; Piatetsky-Shapiro and Frawley, 1991). Knowledge discovery in databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases (Frawley *et al.*, 1991). Data mining represents the integration of several fields, including machine learning, database systems, data visualization, statistics, and information theory.

The concept of *knowledge discovery* has of late gained the attention of the business community. One main reason for this is the general recognition of the need to 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 the knowledge discovery process. In this work, the extracted knowledge is represented as a set of cases or records that can be formally defined as a relationship between a set of attributes and a decision.

The main task carried out in this research is classification, which is the process of ascertaining common properties among different objects and classifying the objects into classes.

Classifying particular situations and/or events as belonging to a certain class is probably the most common data mining problem faced by people in real life applications. Diagnosing diseases, predicting stock market evolution, profiling higher-priced houses or assessing risk in insurance policies are all examples that can be viewed as classification problems. In order to solve these problems, accurate classifier systems or models must be built.

The problem of classification has been widely studied by researchers in the artificial intelligence (AI) field. Several computational intelligence methodologies have been applied to construct such a classifier from particular cases or data. Difficulties include how to represent and work with different types of data, dealing with missing or unknown values and ensuring efficiency.

The database community focuses on searching for effective and efficient classification algorithms. Their work involves either developing new and efficient classification algorithms or further advancing the existing AI techniques, for example extracting rules in "if ... then ..." form that can be applied to large databases (Agrawal *et al.*, 1992; 1993(*a*); 1993(*b*); Ling and Zhang, 2002).

As real life classification applications usually have several features, it increases the complexity of the classification task. It is common for a class label of an object to

depend only on the values of a few features. Knowledge extraction in the classification context is the process of selecting the most important features or attributes from the information systems or a dataset (Shiba *et al.*, 2003(*c*)). Choosing a subset of the features may increase accuracy and reduce complexity of the acquired knowledge.

Selecting an optimal set of features for a given task is a problem which plays an important role in a wide variety of contexts including pattern recognition, adaptive control, and machine learning. Our experience with traditional feature selection algorithms in the domain of machine learning has led to an appreciation of their computational efficiency and a concern for their brittleness. In this research, the features selection task is based on slicing. This new approach assists in identifying the subset of features used in computing the similarity measures needed by classification algorithms.

1.2 Problem Statement

Analysing and mining a real or artificially generated large database are well-known problems in data mining. In managing large databases that may contain thousands of cases or objects and large attribute or features size, most machine learning algorithms extract massive amount of knowledge in the form of decision tree, rules or set of weight in neural network. The case classification is the most important task in managing this large database. All classification approaches depend critically on the availability of a predefined set of classes or categories that may be used to classify the cases. The main problem to solve then is to select the most appropriate class for the problem situation under examination.

This is a classical problem. It is therefore not surprising that different solutions have emerged based on distinct computational intelligence methodologies such as the Statistical Approach, Case-Based Reasoning, Evolutionary Computation, Neural Networks and Fuzzy Logic. At an abstract level, they all produce a set of rules and it will be interesting to compare their performance and analyse the possibilities for hybridization.

The performance of most practical classifiers improved when correlated or irrelevant features of case are removed (Dong and Kothari, 2003). Based on this fact, and the previous classification accuracy results obtained by other researchers, which are not good enough, it is interesting to investigate the optimal way to improve the classification accuracy. The result of the investigation produces a new

classification approach based on slicing to reduce the number of features that will improve the classification accuracy.

1.3 Objectives of the Research

The main objective of this research is to propose an accurate new feature selection approach based on slicing. The secondary objectives include:

- Improving the discretization equation proposed by Randall and Tony (1996) and extended by Payne and Edwards (1998) to convert continuous attribute values into discrete values.
- Proving that feature selection based slicing can improve classification accuracy.

1.4 Scope of the Research

This study focuses on obtaining new feature selection approach based on slicing technique. The research focuses on using this new slicing technique in data mining especially in the classification task- in essence, how to build a classifier system that is able to classify unseen objectives correctly using the slicing technique. The classification algorithms intend to classify objects better both in terms of accuracy and speed. However, in most of the cases, accuracy is the most important factor to consider (Saykol, 2000; Thamar and Olac, 2002). For this reason, improving the classification accuracy has been set as the major aim of the proposed slicing technique.