BORANG PENGESAHAN STATUS TESIS

JUDUL: An Analysis of Attribute Reduction Techniques For Breast Cancer Data Set

SESI PENGAJIAN: Sesi 2009/2010

Saya WONG HOR YAN mengaku membenarkan tesis ([✓ PSM]/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan penukaran antara institusi pengajian tinggi.
4. Sila tandakan(/)

______	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
______	TERHAD	(Mengandungi maklumat TERHAD yang ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
/	TIDAK TERHAD	

(TANDATANGAN PENULIS)

Alamat Tetap: Lot 192, Lorong 8, Fasa 3A,
Taman Megah, Batu 7, Jalan Labuk,
90000 Sandakan, Sabah, Malaysia.

Tarikh: 30 June 2010

(TANDATANGAN PENYELIA)

Dr. Choo Yun Huoy

Tarikh: 30 June 2010

AN ANALYSIS OF ATTRIBUTE REDUCTION TECHNIQUES FOR BREAST CANCER DATASET

WONG HOR YAN

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Artificial Intelligence)

FACULTY OF INFORMATION COMMUNICATION AND TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2010

DECLARATION

I hereby declare that this project entitled

AN ANALYSIS OF ATTRIBUTE REDUCTION TECHNIQUES FOR BREAST CANCER DATA SET

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : ______________________ Date: 30 June 2010

(WONG HOR YAN)

SUPERVISOR : ______________________ Date: 30 June 2010

(DR. CHOO YUN HUOY)

DEDICATION

To my beloved parents Mr. Wong Wing Hing and Mrs. Ho Siew Ngun and also my brothers.

For giving me all the love and support to pull all this through.

To my supervisor, Dr. Choo Yun Huoy,

For helping me by ensuring that I have been in the correct path all the way.

Thank you very much.

ACKNOWLEDGEMENT

First and the most important, I would like to show my gratitude and appreciation to my supporting supervisor, Dr. Choo Yun Huoy for all her ideas, expert advices, suggestions, and patience in guiding me throughout the project.

Besides, I would also like to thank Puan Kasturi Kanchymalay for her generosity in giving me an opportunity to access her Melaka General Hospital Breast Cancer Data Set collection as part of the project analyses and benchmarking practices.

Also, special gratefulness to Universiti Teknikal Malaysia Melaka for offering this subject, the *Projeck Sarjana Muda* which gives me and also other students a lot of opportunity to gain more knowledge.

Moreover, appreciations are also given to the committee members from the Faculty of Information Communication and Technology of Universiti Teknikal Malaysia Melaka who work very hard in planning for the briefing sessions, talks and also exhibition on seniors' handworks.

Last but not least, I would wish to thanks my friends and course mates for their kindness in sharing their knowledge and resources.

ABSTRACT

Breast cancer is a deadly disease popularly among women but the disease is curable when detected in early stage. However, large number of disease markers in breast cancer data set may affects the quality of prediction. Thus, this project's objectives are to analysis and to benchmark attribute reduction techniques besides developing an attribute reduction tool for breast cancer data set. CRISP-DM is used as the main methodology whereas OOAD is used for the tool development. After the attribute reduction tool is completed, analyses of RELIEF, SVM-RFE and CFS techniques on different data sets are done. Experiments on acquiring classification accuracy are done with Naïve Bayes as the classifier, 10-folds cross validation as the evaluation mode and a random seed of 1 while the ROC values and percentage of reduction are used in comparing the classification performance. The experiments shows that CFS achieved high percentage of reduction and fine ROC values in most experiments conducted while SVM-RFE's performance is considered tolerable although it consume more process time than CFS and RELIEF. The experiments also show that RELIEF bore exceptional results for the Wisconsin Breast Cancer data. Thus RELIEF is suggested for numeric-valued attributes and large or artificial data sets while SVM-RFE is good for data with mostly nominal-valued attributes, real-world data with more training data and less testing data. Then, as CFS performs excellently it is recommended for processing numeric-valued attributes and real-world data sets. Future recommendation will be comparing more techniques with more different data set.

ABSTRAK

Kanser payudara merupakan antara penyakit yang membawa maut khasnya di kalangan wanita tetapi penyakit ini masih boleh dirawat sekiranya ditemui pada peringkat awal. Namun demikian, dengan semakin meningkatnya jumlah maklumat yang disimpan dalam pangkalan data, ketepatan ramalan berkemungkinan besar akan terjejas. Maka, objektif projek ini adalah untuk menganalisa dan membangunkan sebuah aplikasi khas untuk proses *attribute reduction* bagi penyakit kanser payudara bagi mengatasi masalah di atas. Methodologi yang bername *CRISP-DM* akan diguna sebagai methodology utama bagi projek ini manakala methodology *OOAD* akan digunakan untuk pembangunan aplikasi. Analisa ke atas *RELIEF, SVM-RFE* dan *CFS* akan dijalankan sejurus siapnya aplikasi *attribute reduction* tersebut. Eksperimen menggunakan *Naïve Bayes classifier* dijalankan bersama *10-folds cross validation* dan *random seed of 1* manakala *ROC values* dan *percentage of reduction* digunakan sebagai cara penilaian ke atas keputusan yang diperolehi. Eksperimen menunjukkan *CFS* memperolehi keputusan yang baik dalam manakala *SVM-RFE* hanya mendapat keputusan yang sederhana dan *RELIEF* menunjukkan keputusan yang bagus untuk *Wisconsin Breast Cancer Data Set*. Maka *RELIEF* dicadangkan untuk data yang bersaoz besar dengan *numeric-valued attributes*, SVM-RFE untuk *more training-real-world* data dengan *nominal-valued attributes* dan *CFS* dicadangkan untuk memproses *numeric-valued attributes* dan *real-world data.* Cadangan untuk penambahbaikan projek ini dalam masa hadapan adalah membuat perbandingan dengan lebih banyak teknik yang lain dan dengan pangkalan data yang lain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NO	ABBREVIATIONS	NAME
1	RELIEF	Relief-F
2	SVM-RFE	Support Vector Machine with Recursive Feature Elimination
3	CFS	Correlation-based Feature Selection
4	CRISP-DM	Cross Industry Standard Process for Data Mining
5	OOAD	Object-Oriented Analysis and Design

LIST OF ATTACHMENTS

LIST OF EQUATIONS

CHAPTER I

INTRODUCTION

1.1 Project Background

Sometimes too much information in a certain data set can reduce the effectiveness of data mining. Some of the columns of data attributes assembled for building and testing a model may not contribute meaningful information to the model. Some may actually detract from the quality and accuracy of the model. Irrelevant attributes simply add noises to the data and affect model accuracy. These noises may increase the size of the model and even increase the time and system resources needed for model building and scoring. Moreover, data sets with many attributes may contain groups of attributes that are correlated. These attributes may actually be measuring the same underlying feature. Their presence together in the build data can skew the logic of the algorithm and affect the accuracy of the model. Wide data or data with a lot of attributes generally presents processing challenges for most of the data mining algorithms. Model attributes are the dimensions of the processing space used by the algorithm. The higher the dimensionality of the processing space, the higher the computation cost will be in algorithmic processing. To minimize the effects of noise, correlation, and high dimensionality, some form of dimension reduction is sometimes a desirable preprocessing step for data mining. Thus, to overcome this problem, attribute reduction techniques are used.

1.2 Problem Statements

Breast cancer data sets are usually consists of quite a number of attributes and instances which make predictions may not as accurate as it might be able to be. Thus, attribute reduction are useful to point out those attributes which contribute the most in term of accuracy. However, there are scores of attribute reduction techniques available with no clear guidance on attribute reduction techniques to be used on breast cancer data sets.

Thus, to deal with this problem, a suggestion on benchmarking suitable attribute reduction techniques for breast cancer data sets is initiated while an attribute reduction tool is developed to assist on the benchmarking progress.

1.3 Project Objectives

The objectives for developing an attribute reduction tool as suggested are as follows:

i. To benchmark selected attribute reduction techniques on breast cancer data set.

ii. To develop an attribute reduction tool with graphical user interface which to run the experiments.

1.4 Project Scopes

i. Modules to be developed and performed:

- A module to enable users to load different data sets.
- A module to run the attribute reduction techniques.
- A module to display the results for every cross validation folds and a subset of best attributes as the final result.
- A module to enable users to save results.
- A module to analysis the results for different attribute reduction techniques.
- A module for benchmarking suitable attribute reduction techniques in term of breast cancer data set.

ii. Limitation:

- Only 3 sets of breast data sets are used which is the Wisconsin Breast Cancer data set, Risk Model data set and the Melaka General Hospital Breast Cancer data set.
- Only 3 types of attribute reduction techniques will be used during the development of the proposed tool which is the Relief F (RELIEF) technique, the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) technique and the Correlation-based Feature Selection (CFS) technique.
- This tool only focuses on the attribute reduction process.
- This tool only accepts CSV and ARFF data set files.
- Benchmarking uses 10-folds cross validation as the evaluation mode and a random seed of 1.
- Benchmarking uses ROC as the performance measure.
- Benchmarking uses a threshold of 0 for CFS technique (Brank et al (2002)) (Morariu, Vintan and Tresp (2006)) and a threshold of average merit for RELIEF and SVM-RFE techniques.