PREDICTING ACCURACY OF INCOME A YEAR USING ROUGH SET THEORY

ZURAIHAH BINTI NGADENGON

UNIVERSITI UTARA MALAYSIA

2009

PREDICTING ACCURACY OF INCOME A YEAR USING ROUGH SET THEORY

A thesis is submitted to the College of Arts and Sciences in partial fulfilment of the requirement for the degree

Master of Science (Intelligent System)

Universiti Utara Malaysia

By

Zuraihah Binti Ngadengon

KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

ZURAIHAH NGADENGON
(89714)

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

PREDICTING ACCURACY OF INCOME A YEAR USING ROUGH SET THEORY

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **DR. SHAIDAH JUSOH**

Tandatangan
(Signature) : _____ Tarikh (Date) : 18/11/2009

Nama Penyelia Kedua
(Name of 2nd Supervisor): **MISS ANIZA MOHAMED DIN**

Tandatangan
(Signature) : _____ Tarikh (Date) : 18/11/2009

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of College of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts there of for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean of College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman.

ABSTRAK

Objektif utama kajian ini dilaksanakan untuk mengukur ketepatan set data *Adult,* samada pendapatan yang diperolehi oleh responden-responden ini melebihi $50 ribu setahun atau tidak. Selain itu, objektif yang hendak dicapai adalah untuk memperolehi kaedah terbaik bagi *discretization, reduction, split factor* dan juga *classifier* di mana kaedah-kaedah ini akan digunakan untuk membangunkan sebuah model klasifikasi (*classification model*). Kajian ini menggunakan pendekatan *rough set* dan perisian *Rosetta* untuk mengukur ketepatan data, manakala *Knowledge Data Discovery* (KDD) digunakan sebagai metodologi bagi menjadi garis panduan perlaksanaan kajian. Kajian ini hanya menggunakan 24, 999 data daripada keseluruhan 48, 842 data. Data ini kemudiannya, dibahagikan kepada dua bahagian, iaitu sebagai data latihan (*training data*) dan juga data pengujian (*testing data*). Pembahagian ini dilakukan dengan menggunakan sembilan faktor pembahagian (*split factor*), iaitu 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 dan 0.9. Keputusan yang diperolehi menunjukkan kaedah terbaik untuk *discretization* ialah *Naïve Algorithm*, *split factor* ialah 0.6, *reduction* ialah *Johnson's Algorithm* dan *classifier* ialah *Standard Voting*. Berdasarkan kepada keputusan yang diperolehi peratus ketepatan tertinggi untuk model spesifikasi yang dibina adalah 87.12%. Ini menunjukkan, pendekatan *rough set* sangat berguna untuk menganalisis set data *Adult* kerana ketepatan yang diperolehi adalah lebih baik berbanding kaedah yang digunakan sebelumnya.

ABSTRACT

The main objective of the experiments is to predict the accuracy of Adult dataset whether the income exceeds $50K per year or below $50K. Specifically, the objectives are to determine the best discretization method, split factor, reduction method, classifier and to build the classification model. In the experiments, the prediction of accuracy of the Adult dataset is developed by using rough set theory and Rosetta software while Knowledge Data Discovery (KDD) is used as the methodology. The Adult dataset that had been used in the experiments is comprises of 48, 842 instances but only 24, 999 instances is used along the experiments. Then, the data was randomly split into training data and testing data by using nine splits factor, which are 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. The result obtained from the experiments showed that the best discretization method is Naïve Algorithm, the best split factor is 0.6, the best reduction method is Johnson's Algorithm and the best classifier is Standard Voting. The highest percentage of accuracy achieved by the classification model developed using the rough set theory is 87.12%. The experiments showed that rough set theory is a useful approach to analyze the Adult dataset because the accuracy achieved in the experiments exceeds the previous methods that have been used before.

ACKNOWLEDGEMENTS

By the Name of Allah, the Most Gracious and the Most Merciful

I would like to thank my supervisor, Dr Shaidah Binti Jusoh and Miss Aniza Binti Mohamed Din for their contribution and assistance in the development of my project, whether in direct or indirect ways. Thanks also to my parent and family for their support and prayer.

TABLE OF CONTENTS

REFERENCES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligent
KDD	Knowledge Data Discovery

CHAPTER ONE

INTRODUCTION

1.1 Background

Artificial intelligence or AI for short is a computer or program, which exhibit behavior that indicates intelligence. McCarthy (2007) defined AI as a task that involved the use of science and engineering to produce a machine that has the intelligence like a human and it is related to the process of making intelligence computer programs. According to Russell and Norvig (2003), the computer or program can be categorized as intelligent if the system think and act like humans and rationally. This means that, the computer is able to simulate human intelligence in order to solve the problem and this can be done by observing human method. Figure 1 below show the dimension in AI.

The contents of the thesis is for internal user only

REFERENCES

Ahmad, F., Abu Bakar, A. & Hamdan, A. Z., (2009). Extracting interesting financial indicators through rough set approach. *International Journal of Computer Science and Network Security, 9 (9),*189-194.

Awang, M. K. (2000). An evaluation of artificial neural network in predicting the presence of heart disease. *Graduate School, University Utara Malaysia.*

Business Dictionary.com. (n.d). Retrieved July 18, 2009, from the World Wide Web: http://www.businessdictionary.com/definition/quality.html

Breault, J. L. (2002). Data mining diabetic databases: are rough sets a useful addition?. *Tulane University.*

Cabena, P. (1977). *Discovering data mining: from concept to implementation.* New Jersey: Prentice-Hall.

Chen, Y. & Chan, K. (2007). Using data mining techniques and rough set theory for language modeling. *ACM Transactions on Asian Language Information Processing,* 6 (1), 1-19.

DeRosa, M. (2004). *Data mining and data analysis for counterterrorism.* New York: Washington. The CSIS Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence,* 37-54.

Han, J. & Kamber, M. (2001). *Data mining: concepts and techniques.* San Diego, USA: Morgan Kauffman Publishers.

Hahsler, M., Grun, B. & Hornik, K. (2005). arules – a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software October 2005, 14 (15),* 1-25.

Harvey, S. (2007). Mining information from US Census Bureau Data. Computer Science, Aston University, Birmingham, UK.

Hassanien, E. & Ali, M. H. J. (2004). Rough set approach for generation of classification rules of breast cancer data. *INFORMATICA, 15(1),* 23 - 28.

He, H., Jin, H., Chen, J., McAullay, D., Li, J. & Fallon, T. (2006). Analysis of breast feeding data using data mining methods. *AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analystics, 61,* 47 – 52.

Hoontrakul, P. & Sahadev, S. (2006). Application of data mining techniques in the on-line travel industry: A case study from Thailand. *Emerald 26(1),* 60-76.

Hornby, S. A. (2007). *Oxford compact advanced learner's: English-Malay dictionary.* Malaysia: Selangor

Hvidsten, T. R., Bjanger, M. S. White, M. F., Guanglai, B. & Komorowski, J. (1999). Fault diagnosis in rotating machineryusing rough sets and rosetta: extended abstract. Norwegian University of Science and Technology.

I-Chun, H., Nan-Kuang, L. & Shih Hui, T.(2007).Using Adult DB to predict yearly salary: greater or less than 50K in 1994.

Komorowski, J., Polkowski, L., Skowron, A. (1999). *Rough sets: a tutorial.* Rough-Fuzzy Hybridization, 3-98.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *KDD-96 Proceedings, 1996.*

Lin, T.Y., & Cercone, N. (1997). *Rough Sets and Data Mining.* Dordrecht, Boston: Kluwer Academic Publishers.

Lloyd, N. M. , Heffernan, N. T & Ruiz, C. (2007). Predicting student engagement in intelligent tutoring systems using teacher expert knowledge. *Proceedings of the 13th International Conference of Artificial Intelligence in Education. Marina del Rey, CA. USA. July 2007.* pp. 40 – 49.

Mary, R. L. (1999). The gender impact of temporary virtual work groups. *IEEE Transactions on Professional Communication, 42 (4),* 276 – 285.

Ning, S., Ziarko, Hamilton, W, J.& Cercone, N.(1995). Using rough sets as tools for knowledge discovery. *KDD'95 Proceedings First International Conference on Knowledge Discovery Data Mining. Montreal, Que., Canada, AAAI.* pp. 263–268.

Mccarthy , J., Minsky, M. L., Rochester, N. & Shannon, C.E. (1995). A proposal for the Dartmouth summer research project on artificial intelligence.

Øhrn, A. & J. Komorowski (1997). ROSETTA: A Rough Set Toolkit for Analysis of Data. *Duke University Press,* 403-407.

Olson, L. D. & Delen, D. (2008). *Advanced data mining techniques.* Springer, Berlin.
Pawlak, Z. (1982). Rough sets. *International Journal Computer and Science, 11,* 341-356.

Russel, S.J & Norvig, P (1999). Artificial intelligence: a modern approach. Prentice Hall: New Jersey.

Ruzgar, N. S., Unsal, F. & Ruzgar, B. (2008). Predicting business failures using the rough set theory approach: The case of the Turkish banks. *International Journal Of Mathematical Models And Methods In Applied Sciences , 1 (2),* 57 - 64.

Seifert, J. W. (2004). Data mining an overview. *Congressional Research Servives, 2004.*

Sifer, M. (2001). Querying web site visitor trend data with coordinated nested bar and pie charts. *Proceedings of the Pan-Sydney Area Workshop on Visual information Processing.* May 2001.

Saygin, A.P, Cicekli, I & Akman, V. (2000). Turing Test: 50 years later. *Minds and Machines, 10,* 463–518.

Tay F.E.H. & Shen L. (2002). Economic and financial prediction using rough sets model. *European Journal of Operational Research,* 141.

The Borneo Post Online. (2009, February 2). Mampukah miskin sifar dicapai menjelang 2010?. Retrieved July 18, 2009, from the World Wide Web: http://www.theborneopost.com/?p=47143.

Turing. A. M. (1950). Computing Machinery And Intelligence. *Computing Machinery And Intelligence Mind, 59,* 433-460.

Wang, H. & Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. *Emerald 108(5),* 622-634.

Zadrozny, B. (2004). Learning and evaluating classiers under sample selection bias. *Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.*

Ziarko, W. (1999). Discovery through rough set theory. Knowledge Discovery: viewing wisdom from all perspectives. *Communications of the ACM, 42* (11).