DATA MINING CLASSIFICATION TECHNIQUES AND PERFORMANCES ON MEDICAL DATA

A thesis submitted to the Faculty of information Technology in partial fulfillment of the requirement for the degree

Master of Science (Intelligent System),

Universiti Utara Malaysia,

By

YAHYIA MOHAMMED M. ALI BENYEHMAD

PUSAT PENGAJIAN SISWAZAH
(Centre for Graduate Studies)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

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

YAHYIA MOHAMMED M. ALI

calon untuk Ijazah
(candidate for the degree of) **MSc. (Int. Sys.)**

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

DATA MINING CLASSIFICATION TECHNIQUES AND PERFORMANCE ON MEDICAL DATA

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 filed is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **DR. FAUDZIAH AHMAD**

Tandatangan
(Signature) : __________ Tarikh (Date): 2/10/06

Nama Penyelia Kedua
(Name of 2nd Supervisor): **DR. SHAIDAH JUSOH**

Tandatangan
(Signature) : __________ Tarikh (Date): 2/10/06

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a post graduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or part, for scholarly purpose may be granted by my supervisor(s) or, in thesis absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without any 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.

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

Dean of Faculty of Information Technology
Department of Computer Science
Universiti Utara Malaysia
06010 Sintok
Kedah Darul Aman

ABSTRACT

This study evaluates the performance of classification techniques with the application of several software, among them are Rosetta, Tanagra, Weka and Orange. The classification technique has been tested on six medical datasets from the UCI Machine Learning Repository. The study will help researchers to select the best suitable technique of classification problem for medical datasets in term of classification accuracy. In this thesis, sixteen classification techniques have been evaluated and compared. These are Radial Basis Function (RBF), Multilayer Perceptron (MLP) Neural Networks, Multi Linear Regression (MLR), Logistic Regression (LR), Classification Tree (ID3, C4.5, J48, CART), Naive Bayes (NB), Support Vector Machines (SVM), k- Nearest Neighbors (kNN), Linear discriminate analysis (LDA), Rule based classifier, Standard voting, Voting with object tracking and Standard/ tuned voting (RSES). The experiments have been validated using 10-fold cross validation method. The results of the study shows that the most suitable classification technique is NB with an average classification accuracy of 90.13% and an average error rate of 9.87%. The worst classification technique is SLR with an average classification accuracy of 50.16% and an average error rate of 49.84%. The classification techniques has been ranked from the best to the worst based on average classification accuracy and average error rate. The top of the rank is NB and the bottom is SLR. The sequence of ranking from the best to the worst is NB, LDA, LR, SVM, C4.5, MLP, RBF, kNN, RuleB, ID3, CART, J48, SV, RSES, V, and SLR.

ACKNOWLEDGEMENT

Praise to Allah S.W.T whose blessing and guidance have helped me through entire project works. Peace be upon our Prophet Mohammad S.A.W, who has given light to mankind. My most sincere appreciation goes to my beloved parents Mohammed M. Ali BenYehmad and Tata Yahia for their patience, prayers and understanding over entire period of my study, although I was away from home but their care and concerned never make me felt alone. Also for my family who always there that gave me love and encourage me along the way.

My sincere gratitude and deep appreciation to my supervisors, Dr. Faudziah Bt Ahmad and Dr. Shaidah Jusoh, Faculty of Information Technology, Universiti Utara Malaysia (UUM) I have learned much about data mining classification techniques. I wish to acknowledge his assistance and time, provided excellent facilities, support and guidance throughout the project also for his advice during this project. Last but not least, a special thanks to my dear friends for their encouragement throughout the study. There is a tremendous sense of achievement in completing this study. To thanks to all the lecturers and members of MSc. Intelligent System batch September 2006, all the best.

Yahia Mohammed M Ali BenYehmad
Faculty of Information Technology
Department of Computer Science
Universiti Utara Malaysia
September 2006

TABLE OF CONTENTS

Chapter 1: Introduction

Chapter 2: Literature Review

Chapter 3: Methodology

Chapter 4: Finding and Discussion

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 *

INTRODUCTION

1.1 Overview

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful means for analysis, interpretation and extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans (Frawley *et al.*, 1992).

The kinds of patterns that can be discovered depend upon the data mining tasks employed. Generally, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. There are many ways to approach data mining problems, including creating statistical models, classification, predictive modeling, clustering, finding association rules and sequence analysis, and anomaly detection. Classification is one of the most important tasks in data mining where its main

REFERENCE

Antonie, M.-L., Za¨ÿane, O.R. and Coman, A. (2001)." Application of Data Mining Techniques for Medical Image Classification". *Proceeding of the Second International Workshop on Multimedia Data Mining (MDM/KDD).*San Francisco, USA, August.

Apte, C., and Hong, S. J. 1996. Predicting Equity Returns from Securities Data with Minimal Rule Generation. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 514–560. Menlo Park, Calif.: AAAI Press.

Begg, R. and Kamruzzaman. (2003). "A Comparison of Neural NetFvorks and Support Vector Machines for Recognizing Young-Old Gait Patterns". *IEEE Transactions on Biomedical Engineering*, Vol.1, pp.354-358.

Berry, M.J.A. and Linoff, G. (1997). "Data Mining Techniques for Marketing, Sales and Customer Support". New York: Wiley.

Brazdil, P., Gama, J., and B.Henry (1994). "Characterizing the Applicability of Classification Algorithms Using Meta-Level Learning". *In Proc. of the European Conference of Machine Learning.*

Calvo, R. A. and. Ceccatto, H. A. (2000). "Intelligent document classification" . *Intelligent Data Analysis*, 4(5).

Dunham, M.H. (2003), "*Data Mining Introductory and Advanced Topics*", 1st Edition Pearson Education (Singaphore) Pte. Ltdl.

Dudani, S. (1975). The distance-weighted k-nearest-neighbour rule. *IEEE Transactions on Systems,Man and Cybernetics*, SMC-6(4):325Œ327.

Djorgovski, S. G., Fayyad, U. M. and Weir, N. (1996). From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey. *AI Magazine* 17(2): 51–66.

Demsar, J., Zupan, B., Leban G. (2004). 'Orange: From Experimental Machine Learning to Interactive Data Mining', White Paper (www.ailab.si/orange), *Faculty of Computer and Information Science, University of Ljubljana.*

Fayyad, U., Shapiro, P. and Smyth, P. (1996). "From Data Mining to Knowledge Discovery: An Overview," *Advances in Knowledge Discovery and Data Mining*, AAAI press/The MIT press, Menlo Park, CA, pp.1-34.

Frawley, W.J. and Piatetsky-Shapiro, G. (1992). "Knowledge Discovery in Databases". AAAI/MIT Press.

Gahegan, M., German, G.W.H. and West, G. (2000). "Statistical and AI Techniques in GIS Classification: A Comparison". School of Computing, Curtin University of Technology, Bentley, Western Australia 6102. 2 Dept Geography, Penn State, University Park, PA 16802 USA.

Güven, A., Kara, S. and Okandan, M. (2003) "Application of artificial neural networks in the pattern elektroretinographical diagnosis of eye diseases", *International XII. Turkish*

Symposium on Artificial Intelligence and Neural Networks (TAINN), Çanakkale, Turkey.

Giannotti, F., Manco, G. and Franco Turini, F. (2004). "Towards a Logic Query Language for Data Mining". *Database Support for Data Mining Applications*, LNAI 2682,pp.76 – 94.

Huang, Y.-L., Wang, K-L. and Chen, D.-R. (2005)." Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines". *Neural Computing & Applications* , Issue: Volume 15, Number 2. 15:164-169.

Houston, A.L., Chen, H., Hubbard, S.M., Schatz, B.R., Ng, T.D., Sewell, R.R. and Tolle, K.M. (1999). "Medical Data Mining on the Internet: Research on a Cancer Information System". *Artificial Intelligence Review* 13: 437–466. Kluwer Academic Publishers.

Han, J. and Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Habrard, A., Bernard, M. and Jacquenet, F. (2003)." Multi-Relational Data Mining in Medical Databases". *Springer-Verlag. EURISE* - Université de Saint-Etienne - 23. 42023 Saint-Etienne cedex 2 – France.

Hussain, F., Liu, H., Tan, C. and .Dash, M. (2002). "Discretization:An Enabling Technique". *Journal of Knowledge Discovery and Data Mining* ,6(4):393 –423.

Kusiak, A., Shah, S. and Dixon, B. (2003). "Data Mining in Predicting Survival of Kidney Dialysis Patients - Invariant object approach". *in Proceedings of Photonics West - Bios, Bass*, L.S. et al. (Eds), *Lasers in Surgery: Advanced Characterization, Therapeutics, and Systems XIII*, Vol. 4949, SPIE, Belingham, WA , pp. 1-8.

Kalousis, A. and Theoharis, T. (1999). "Noemon: Design, implementation and performance results of an intelligent assistant for classifier selection". *Intelligent Data Analysis*, 3(5):319--337.

Lim, T.-S., Loh W.-Y and Shih, Y.-S. (2000). "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classi_cation Algorithms". *Machine Learning*, 40, 203-228.

Lorenz, A., Blüm, M., Ermert, H. And Senge, T. (2000) "Comparison of Different Neuro-Fuzzy Classification Systems for the Detection of Prostate Cancer in Ultrasonic Images". *Bundesministerium für Forschung und Technologic*, D-44780 Bochum, Germany. Grant: 01 KF 8903/2.

Leroy, G. and Rindflesch, T.C. (2004). "Using Symbolic Knowledge in the UMLS to Disambiguate Words in Small Datasets with a Naïve Bayes Classifier". *AMIA Symp.*; p. 381- 385.

Mitchell, T. (1997). "Bayesian Learning, Machine Learning", 154-200. McGraw- Hill.

Merz, C.J. and Murphy, P.M. (1996) .UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California, Irvine, CA. (http:// www .ics.uci.edu/~mlearn/MLRepository.html).

Mitra, P., Mitra, S. and Pal, K. (2002). "Staging of Cervical Cancer with Soft Computing". *IEEE Transa ctions on biomedical engineering*, vol. 47, no.7.

Michie, D., Spiegelhalter, D. J., and Taylor C. C. (1994). "Machine learning, Neural and Statistical classification". New York: *Ellis Horwood.* xiv + 289 pp.

Øhrn, A., Komorowski, J. (1997). "ROSETTA: A Rough Set Toolkit for Analysis of Data". *Proceedings of the Third International Joint Conference on Information Sciences*, Durham, NC, USA, *Department of Electrical and Computer Engineering*, Duke University Vol.3 pp. 403–407.

Øhrn, A. and T. Rowland (2000). "Rough sets: a knowledge discovery technique for multifactorial medical outcomes. *" Am J Phys Med Rehabil 79(1): 100-108.*

Øhrn, A., (1999). "Discernibility and Rough Sets in Medicine: Tools and Applications", PhD thesis, *Department of Computer and Information Science*, Norwegian University of Science and Technology, Trondheim, Norway. NTNU report 1999:133.

Rakotomalala, R. (2005). TANAGRA: "a free software for research and academic purposes", *in Proceedings of EGC'05,* RNTI-E-3, vol. 2, pp.697-702.

Soares, C. and Brazdil, P. B. (2000). "Zoomed Ranking: Selection of Classification Algorithms Based on Relevant Performance Information". *In Proceedings of Principles of Data Mining and Knowledge Discovery*, 4th European Conference (PKDD-2000), 126-135.

Tan, A.C. and GILBERT, D. (2003) "An empirical comparison of supervised machine learning techniques in bioinformatics". *in the Proceedings of the First Asia Pacific Bioinformatics Conference*, Vol. 19.

Todorovski, L. and Dzeroski, S. (1999). "Experiments in meta-level learning with ILP". *In Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-99)*, pages 98--106.

TANAGRA. A Free Data Mining Software for Research and Education. (http://eric.univ-lyon2.fr/ricco/tanagra/). 2005.

Xiong, L., Chitti, S. and Liu, L. (2004). "Mining Multiple Private Databases using a Privacy Preserving kNN Classifier". *Submitted. Available as technical report, College of Computing, Georgia Institute of Technology.*

Yang, Y. and Liu, X. (1999) "A re-examination of text categorization methods". *In 22nd Annual International SIGIR,* pages 42–49, Berkley.

Zheng, J., Yan, H., Jiang, Y., Peng, C. and Li, Q. (2003). "Development of a decision support system for heart disease diagnosis using multilayer preceptron". Proc. IEEE, Page(s): V-709 - V-712 vol.5.