AN EXPERIMENTAL STUDY OF CLASSIFICATION ALGORITHMS TRAINING PERFORMANC

A thesis submitted to the Faculty of Information Technology in partial

Fulfillment of the requirements for the degree

Master of Science (Intelligent System)

Universitiy Utara Malaysia

By

Khald Ali I. Aboalayon

JABATAN HAL EHWAL AKADEMIK
(Department of Academic Affairs)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

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

KHALD ALI I. ABOALAYON

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)

AN EXPERIMENTAL STUDY OF CLASSIFICATION ALGORITHMS TRAINING PERFORMANCE

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): **MR. AZIZI AB. AZIZ**

Tandatangan
(Signature) :

AZIZI BIN AB. AZIZ
Pensyarah
AISIG Jabatan Sains Komputer
Fakulti Teknologi Maklumat
Universiti Utara Malaysia

Tarikh
(Date) : 10/07/05

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a post graduate degree from University 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 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 University 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
University Utara Malaysia
06010 Sintok
Kedah Darul Aman

ABSTRACT

This thesis evaluates the training performance of classifiers in terms of Root Mean Square Error (RMSE), Training Time and Complexity. The study was based on different data set that were obtained from UCI machine learning database and tested by the WEKA software machine learning tools. The aim of this study is to experiment several classifiers with different data sets to find out the best classifier for a certain data set like nominal, numerical and both, according to the objective of this research.

ACKNOWLEDGEMENT

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

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 Ali Aboalayon and Fatmah Ali 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 supervisor, Mr. Azizi Bin Ab Aziz, Faculty of Information Technology, Universiti Utara Malaysia (UUM) I have learned much about classification algorithm 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 July 2005, all the best.

Khald Ali I. Aboalayon
Faculty of Information Technology
Department of Computer Science
University Utara Malaysia

July 2005

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: METHODOLOGY

CHAPTER 4: RESULTS

CHAPTER 5: CONCLUSION... 48

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This section describes the context of the study that includes the introduction to classification techniques, followed by the problem statement, the objectives of the study, the scope of study and finally, the significance of the study.

1.1 Classification

Classification is one of the data mining techniques. Classification maps data into predefined groups or classes. It is often referred as supervised learning because the classes are determined before examining the data. Classification algorithms require the classes to be defined based on data attribute values. Figure 1.1 shows classification tasks.

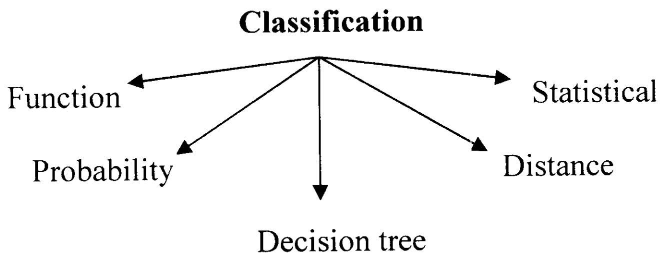

Figure 1.1: Classification algorithm categorization

One common classification scheme based on the use of distance measures is K Nearest Neighbours (KNN). The KNN technique assumes that the entire set includes not only the data in the set, but also the desired classification for each item (Dunham, 2003).Whereby,

REFERENCE

A.AbuBakar, Sulaiman, A., & M.Selamt. (2001). An Improved Rovsh Classification Model: A comparison with Neural Classifier. *Journal of Hstihk of Maths and Comp. Scienes,, Vol 12,1), pp. 43.*

Begg, R., & Kamruzzaman, J. (2003). *A Comparison of Neural Networks and Support Vector Machines for Recognizing Young-Old Gait Patterns.*

Bloom, J. Z. (2003). *Tourist market segmentation with linear and non-linear techniques.* South Africa.

Bojarczuk, C. C., Lopes, H. S., & Freitas, A. A. (2003). *An innovative application of a constrained-syntax genetic programming system to the problem of predicting survival of patients.*

Buntine, W. (1991). *Introduction to IND and Recursive Partitioning, .*

Calvo, R., Lee, M., & Li, X. (2000). *Managing content with automatic document classification.*

Dunham, M. H. (2003). *Data Mining Introductory and Advanced Topics.* New Jersey.

E.W, S., M.J.C, E., & 1, H. J. D. F. (2001). *Application of Shrinkage Techniques in Logistic Regression Analysis.*

Fenga, S., Lib, L., Cena, L., & Huanga, J. (2003). *Using MLP networks to design a production scheduling system.* Old Dominion University,Norfolk, VA 23529, USA.

Forman, G., & Cohen, I. (2004). *Comparison of Classifiers Given Little Training.*

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). *Data Mining in Bioinformatics using Weka.*

Gay, L. R., & Diebl, P. L. (1994). *Research Methods For Business and Management.* Miami, Florida.

German, West, & Gahegan. (2000). *Statistical and AI Techniques in GIS Classification: A Comparison.*

Gonalves, T., & Quaresma, P. (2004). *A preliminary approach to the multilabel classication problem of Portuguese juridical documents.*

GVEN, A. e., Karak, S., & Okandan, M. (2003). Application of Artifical Neural Network in The Pattern Elektroretinographical Diagnosis of Eye Diseases. *International XII. Turkish Symposium on Artificial Intelligence and Neural Networks*(Erciyes University, Civil Aviation School, Kayseri, 38039, Turkey.).

I.Florea, F., Rogozan, A., Bensrhair, A., & Darmoni, S. e. J. (2004). *Comparison of Feature-Selection and Classification Techniques for Medical Images Modality Categorization.*

J. Shavlik, Mooney, R., & Towell, G. (1995). *Symbolic and Neural Network Learning Algorithms: An Experimental Comparison.*

Kaski, S. (1997). *Data Exploration Using Self Organizing Mapping.*

Khoussainov, R., Zuo, X., & Kushmerick, N. (2002). *Grid-enabled Weka: a toolkit for machine learning on the Grid.*

Kononenko, I. (1992). *Inductive and Bayesian Learning in Medical Diagnosis.*

Koutroumbas, K., Paliouras, G., Karkaletsis, V., & Spyropoulos, C. D. (2001). *Comparison of Computational Learning Methods on a Diagnostic. Cytological Application.*

Kucukyilmaz, A. (2005). *Pattern Classification: A Survey and Comparison.*

Land, F. W. (1997). *Stock Price Prediction using Neural Networks.*

Larson, R. R. (2002). *A Logistic Regression Approach to Distributed IR.*

Loh, W. Y., & Shih, Y. S. (2000). *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classication Algorithms.* Paper presented at the Machine Learning,.

Malerbo, Appice, Bellino, Ceci, & Pallota. (2001). Stepwise Induction of Model Trees, Springer Verlag.

Mehmet, Tolun, & Abu-Soud, S. (1999). *An Inductive Learning Algorithm for Production Rule Discovery.*

Mendro, R. L., Jordan, H. R., Gomez, E., Anderson, M. C., & Bembry, K. L. (1998). *An Application of Multiple Linear Regression in Determining Longitudinal Teacher Effectiveness.* Dallas, Texas.

Murphy, & Aha, D. W. (1994). *UCI Repository of Machine Learning Databases.*

Pal, M., & Mather, P. M. (2003). *Support Vector classifiers for Land Cover Classification.*

Polikar, R., L.Udpa, S. S. U., & Honavar, V. (2000). *LEARN++: Annincrement Learning Algorithm for Multilayer Perceptron Networks.* Ames: Dept. of Electrical and Computer Engineering,.

Qian, J. (2003). *Application of Logistic Regression in Analysis of e-rater Data.*

Soman, T., & O.Obobbie, P. (2005). *Classification of Arrhythmia Using Machine Learning Techniques1,2.*

Tan, A. C., & Gilbert, D. (2003). *An empirical comparison of supervised machine learning techniques in bioinformatics.*

Tylor, & Spiegelhater, M. (1994). *Machine Learning, Neural and Statistical Classification.*

Witten, I. H., & Frank, E. (2000). *Data Mining: practical Machine Learning Tools.* New York.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (2000). *Weka: Practical Machine Learning Tools and Techniques with Java Implementations.*

Yan, H., Zheng, J., Jiang, Y., Pengl, C., & Li, Q. (2003). *Development of A decision Support System For Heart Diseas Diagnosis Using MUL Multilayer Perceptron.* Vegas.

Yoachims, T. (1998). *Text Categorization with support Machine.* Paper presented at the Proceeding of ECML European Conference on Machine Learning.

Yuan, X., Yuan, X., Buckles, B. P., & Zhang, J. (2003). *A Comparison Study of Decision Tree and SVM to Classify Gene Sequence.*

Zenko, B., Todorovski, L. c., & D‹zeroski, S. s. (2001). *A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods.*