Investigating the Impact of Different Representations of Data on Neural Network and Regression

This thesis is presented to the Graduate School
In fulfilment of the requirements for
Master of Science (Information Technology)
Universiti Utara Malaysia

By

Ehab A. Omer El Fallah

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)

EHAB A. OMER EL FALLAH

calon untuk Ijazah
(candidate for the degree of) **MSc. (Information Technology)**

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

INVESTIGATING THE IMPACT OF DIFFERENT REPRESENTATION OF DATA ON NEURAL NETWORK AND REGRESSION

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): **ASSOC. PROF. FADZILAH SIRAJ**

Tandatangan
(Signature) : Fadzilah

Tarikh
(Date) : 22/06/08

PERMISION TO USE

This thesis presents a partial fulfilment of the requirement for a postgraduate degree from Universiti Utara Malaysia. I agree that the Universiti Library may make it freely available for inspection. I further agree that the permission for copyright of this these in many manners, in whole or part, for scholarly purpose may be granted by my supervisor or, in their absence by the Dean of the Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood 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 Graduate School
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman.

ABSTRACT

In this research the impact of different data representation on the performance of neural network and regression was investigated on different datasets that has binary or Boolean class target. In addition, the performance of particular predictive data mining model could be affected with the change of data representation. The seven data representations that have been used in this research are As_Is, Min Max normalization, standard deviation normalization, sigmoidal normalization, thermometer representation, flag representation and simple binary representation. Moreover, all data representations have been applied on two datasets which are Wisconsin breast cancer and German credit dataset. As a result, the neural network performance is better than logistic regression on both datasets if we exclude the thermometer and flag representations. For datasets having a binary or Boolean target class, flag or thermometer binary representation is recommended to be used if logistic regression analysis is performed. Meanwhile, As_is representation, min max normalization, standard deviation normalization or sigmoidal normalization is recommended for neural network analysis on datasets having binary or Boolean target class.

TABLE OF CONTENTS

LIST OF TABLES

Title Page

CHAPTER 1

INTRODUCTION

1.0 Background

Data mining has been used widely in many different areas and domains to extract useful information from large amounts of data (Leung *et al.*, 2001). Instead of being its own field, data mining is a combination of several fields such as computer science, artificial intelligence and statistics (Remondino & Correndo, 2005). In addition, data mining is a crucial step in the Knowledge Discovery in Databases (KDD) process which comprises of data cleaning, data consolidation, data selection, data transformation, data mining, pattern analysis and knowledge presentation (Ozekes & Osman, 2003). There are two types of data mining models, namely the predictive and the descriptive (Kusiak, 2006; Remondino & Correndo, 2005; Ozekes & Osman, 2003). Descriptive data mining aims to summarize data and extract interesting properties from the data, while predictive data mining aims to build models and predict future behaviours. Data mining tasks can be grouped into four categories which are association, summarization, classification, clustering and trend analysis (Luo, 2008). There are many different methods of predictive data mining, for example, prediction, classification, regression, and time series.

The contents of the thesis is for internal user only

7. REFERENCES

Abraham, R., Simha, J. B. & Iyengar S. S (2007). Medical Datamining with a New Algorithm for Feature Selection and Naive Bayesian Classifier. *10th International Conference, ICIT 2007: Information Technology, 17-20 Dec 2007* (pp. 44-49).

Altun, H., Talcinoz, T. & Tezekiei B. S. (2000). Improvement in the Learning Process as a Function of Distribution Characteristics of Binary Data Set. 10th Mediterranean Electrotechnical Conference, 2000,Vol. 2 (pp. 567-569).

An, L. & Tong, L. (2005). A rough neural expert system for medical diagnosis. Proceeding of International Conference, ICSSSM '05: Services Systems and Services Management, 13-15 June 2005, Vol. 2 (pp.1130-1135).

Ashkenazy, Y., Ivanov, P. C., Havlin, S., Peng, C. K., Yamamoto, Y., Goldberger, A. L. & Stanley, H. E. (2000). Decomposition of Heartbeat Time Series: Scaling Analysis of the Sign Sequence. Computers in Cardiology 2000, 27, 139-142.

Bengtsson, B., Bizios, D. & Heijl, A. (2005). Effects of Input Data on the Performance of a Neural Network in Distinguishing Normal and Glaucomatous Visual Fields. Investigative Ophthalmology and Visual Science, 46, 3730-3736.
(URL: http://www.iovs.org/cgi/content/abstract/46/10/3730)

Burke, H. B., Rosen, D. B. & Goodman, P. H. (1994). Comparing Artificial Neural Networks to Other Statistical Methods for Medical Outcome Prediction. IEEE International Conference: Neural Networks 1994, IEEE World Congress on Computational Intelligence, 27 June – 2 July 1994, Vol. 4 (pp. 2213-2216).

Chen L. & Kamel M. S. (2007). A New Design of Multiple Classifier System and its Application to the Classification of Time Series Data. IEEE International Conference, ISIC 2007: Systems, Man and Cybernetics, 7-10 October 2007 (pp. 385-391).

Delen, D. & Patil, N. (2006). Knowledge Extraction from Prostate Cancer Data. Proceedings of the 39th Annual Hawaii International Conference, HICSS '06: System Sciences, 04-07 Jan 2006, Vol. 5 (pp. 92b-92b).

El-Sebakhy, E. A., Faisal, K. A., Helmy, T., Azzedin, F. & Al-Suhaim, A. (2006). Evaluation of Breast Cancer Tumor Classification with Unconstrained Functional Networks Classifier. IEEE International Conference: Computer Systems and Applications, 8 March 2006 (pp. 281-287).

Fausett, L. (1994). Fundamentals Of Neural Networks Architectures, Algorithms, and Applications. Upper Saddle River, New Jersey 07458: Prentice Hall.

Forti, M. (2002). Some Extensions of a New Method to Analyze Complete Stability of Neural Networks. IEEE Transactions on Neural Networks, 13(5), 1230-1238.

Halstead, J. B. & Brown, D. E. (2004). Improving Upon Logistic Regression to Predict United States Army Delay Entry Program (DEP) Losses. Proceedings of the 2004 IEEE: Systems and Information Engineering Design Symposium,16-16 April 2004 (pp. 191-201).

Hashemi R. R., Bahar, M., Tyler, A. A. & Young, J. (2002). The Investigation of Mercury Presence in Human Blood: An Extrapolation from Animal Data Using Neural Networks. Proceedings of International Conference: Information Technology: Coding and Computing, 8-10 April 2002 (pp. 512-517).

Hung, C. & Tsai, C. F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. Expert Systems with Applications, 34, 780-787.

Jia, J. & Chua, H. C. (1993). Neural Network Encoding Approach Comparison: An Empirical Study. Proceedings of First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, 24-26 November 1993 (pp 38-41).

Jia, J. & Chua, H. C. (1995). Solving Two-Spiral Problem through Input Data Representation. Proceedings Of IEEE International Conference 1995: Neural Networks, 27 Nov – 1 Dec 1995, Vol. 1 (pp. 132-135). Perth, WA.

Kantardzic, M. (2003). DATA MINING: Concepts, Models, Methods and Algorithms. IEEE Transactions on Neural Networks, 14(2), 464-464.

Kosmelj, K. & Vadnal, K. (2003). Comparison of two generalized logistic regression models: a case study. Proceedings of the 25th International Conference, ITI 2003: Information Technology Interfaces, 16-19 June 2003 (pp. 199-204).

Ksantini, R., Ziou, D., Colin, B., & Dubeau, F. (2008). Weighted Pseudometric Discriminatory Power Improvement Using a Bayesian Logistic Regression Model Based on a Variational Method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2), 253-266.

Kurt, I., Ture, M. & Kurum, A.T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications, 34, 366-374.

Kusiak, A. (2006). Data Mining in Design of Products and Production Systems. Proceedings of INCOM'2006: 12th IFAC/IFIP/IFORS/IEEE Symposium on Control Problems in Manufacturing, May 2006 (pp. 49-53). Saint-Etienne, France.

Lean, Y., Wang, S. & Lai, K.K. (2006). An Integrated Data Preparation Scheme for Neural Network Data Analysis. IEEE Transactions on Knowledge and Data Engineering, 18(2), 217-230.

Lee, A. H. I., Chen, W. C. & Chang, C. J.(2008). A fuzzy AHP and BSC approach for evaluating performance of IT department in the manufacturing industry in Taiwan. Expert Systems with Applications, 34, 96-107.

Leung Y., Ma J. H. & Zhang W. X. (2001). A New Method for Mining Regression Classes in Large Data Sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(1), 5-21.

Lewicke, A., Sazonov, E., Corwin, M. J., Neuman, M., Schuckers, S., & CHIME. (2008). Sleep Versus Wake Classification From Heart Rate Variability Using Computational Intelligence: Consideration of Rejection in Classification Models. IEEE Transactions on Biomedical Engineering, 55 (1), 108-118.

Lin, C. T. & Chen. C. T. (2004). A Fuzzy-Logic-Based Approach for New Product Go/NoGo Decision at the Front End. IEEE Transactions on Systems, Man and Cybernetics, Part A, 34, 132-142.

Lu, C., Devos, A., Suykens, J. A. K., Arus, C. & Huffel, S. V. (2007). Bagging Linear Sparse Bayesian Learning Models for Variable Selection in Cancer Diagnosis, IEEE Transactions on Information Technology in Biomedicine, 11(3), 338-347.

Luo, Q. (2008). Advancing Knowledge Discovery and Data Mining. International Workshop on Knowledge Discovery and Data Mining, 23-24 January 2008 (pp. 3-5).

Nawi, N. M., Ransing, M. R. and Ransing R. S. (2006). An Improved Learning Algorithm Based on The Broyden-Fletcher-Goldfarb-Shanno (BFGS) Method For Back Propagation Neural Networks. Sixth International Conference on Intelligent Systems Design and Applications, October 2006, Vol. 1, pp.152-157.

O'Neal, M.R., Engel, B.A., Ess, D.R. & Frankenberger, J.R. (2002). Neural Network prediction of maize yield using alternative data coding algorithms. Biosystems Engineering, 83, 31-45.

Ortiz-Arroyo, D., Skov, M.K. & Huynh, Q. (2005). Accurate Electricity Load Forecasting with Artificial Neural Networks. International Conference on Computational Intelligence for

Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, 28-30 November 2005, Vol. 1 (pp. 94-99).

Ozekes, S. & Osman, O. (2003). Classification and Prediction in Data Mining with Neural Networks. Istanbul University - Journal of Electrical & Electronic Engineering, 3, 702-712.

Remondino, M. & Correndo, G. (2005). Data Mining Applied to Agent Based Simulation. Proceedings 19th European Conference on Modelling and Simulation, 2005.

Revett, K., Gorunescu, F., Gorunescu, M., El-Darzi, E. & Ene, M. (2005). A Breast Cancer Diagnosis System: A Combined Approach Using Rough Sets and Probabilistic Neural Networks. The International Conference, EUROCON: Computer as a Tool, 2005, Vol. 2 (pp. 1124-1127). Belgrade, Serbia and Montenegro.

Schaefer, G., Nakashima, T., Yokota, Y. & Ishibuchi, H. (2007). Cost-Sensitive Fuzzy Classification for Medical Diagnosis. IEEE Symposium: Computational Intelligence and Bioinformatics and Computational Biology, 1-5 April 2007 (pp. 312-316).

Sewak, M., Vaidya, P., Chan, C. C., & Duan, Z. H. (2007). SVM Approach to Breast Cancer Classification. Second International Multi-Symposiums IMSCCS 2007: Computer and Computational Sciences, 13-15 Aug 2007 (pp. 32-37).

Shen, L. & Tan, E. C. (2005). Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2(2), 166-175.

Shen, A., Tong, R., & Deng, Y. (2007). Application of Classification Models on Credit Card Fraud Detection. International Conference: Service Systems and Service Management, 9-11 June 2007 (pp. 1-4).

Stalbovskaya, V., Ifeachor, E.C., Huffel, S.V., Timmerman, D. (2007). A New Method for Modeling Preoperative Diagnosis of Ovarian Tumors. IEEE Transactions: Biomedical Engineering, 54(11), 2064-2072.

Thomassey, S., Happiette, M. & Castelain, J. M. (2002). Textile items classification for sales forecasting. ESS 2002, fourth European Simulation Symposium and Exhibition Simulation in Industry.

UCI Machine Learning Repository. (n.d.). Retrieved May 15, 2008, from http://www.ics.uci.edu/~mlearn/MLRepository.html.

Wang, W. B., Vundla, S., Sun, H. Y., & Tian, Y. Z. (2007). A Statistical Study on the Relationship between Business Performance Measures and Influencing Factors/Covariates. International Conference, ICMSE 2007: Management Science and Engineering, 20-22 Aug 2007 (pp. 599-613).

Wessels, L.F.A., Reinders, M.J.T., Welsem, T.V. & Nederlof, P.M. (2002). Representation and classification for high-throughput data sets. SPIE-BIOS2002, Biomedial Nanotechnology Architectures and Applications, 4626, 226-237, San Jose, USA, Jan 2002.

Wettayaprasit W. and Sangket U. (2006). Linguistic Knowledge Extraction from Neural Networks Using Maximum Weight and Frequency Data Representation. IEEE Conference: Cybernetics and Intelligent Systems, June 2006 (pp. 1-6). Bangkok.

Xu, L. & Chow, M.Y. (2006). A Classification Approach for Power Distribution Systems Fault Cause Identification. IEEE Transactions on Power Systems, 21(1), 53-60.

Yada, K., Ip, E., & Katoh, N. (2007). Is this brand ephemeral? A multivariate tree-based decision analysis of new product sustainability. ScienceDirect: Decision Support Systems, 44, 223–234

Yang Y. and Chen K. (2006). An Ensemble of Competitive Learning Networks with Different Representations for Temporal Data Clustering. International Joint Conference, IJCNN '06: Neural Networks, 16-21 July 2006 (pp.3120-3127).

Yoon, J., Kwon, Y. S. & Roh, T. H. (2007). Performance Improvement of Bankruptcy Prediction using Credit Card Sales Information of Small & Micro Business. 5th ACIS International Conference: Software Engineering Research, Management & Applications, 20-22 Aug 2007 (pp. 503-512).

Yun, W. H., Kim, D. H., Chi, S. Y. & Yoon, H. S. (2007). Two-dimensional Logistic Regression. 19th IEEE International Conference, ICTAI 2007: Tools with Artificial Intelligence, 29-31 October 2007, Vol. 2 (pp. 349-353).

Zhang, N. & Lu, W.F. (2007). An Efficient Data Preprocessing Method for Mining Customer Survey Data. 5th IEEE International Conference: Industrial Informatics, 23-27 June 2007, Vol. 1 (pp. 573-578). Vienna.

Zhang, D. & Zhou, L. (2004). Discovering Golden Nuggets: Data Mining in Financial Application. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Review, 34(4), 513-522.

Zheng, G., Huang, Y. L., Wang, P. T. & Shu, G. F. (2005). A Study of classification rules on weighted coronary heart disease data. Proceedings of 2005 International Conference: Machine Learning and Cybernetics, 18-21 August 2005, Vol. 3 (pp.1845-1850).

Zhu, D., Premkumar, G., Zhang, X. & Chu, C.H. (2001). Data mining for Network Intrusion Detection: A Comparison of Alternative Methods. *Decision Sciences, 32*(4), 635-660.