EMPIRICAL COMPARISON OF TECHNIQUES FOR HANDLING MISSING VALUES

A thesis submitted to the Faculty of Information Technology in partial
fulfillment of the requirements for the degree
Master of Science (Intelligent System)
Universiti Utara Malaysia

By

Saleh Mansour Mohamed Tikla

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)

SALEH MANSOUR MOHAMED TIKLA

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)

EMPIRICAL COMPARISONS OF TECHNIQUES FOR HANDLING MISSING VALUES

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. WAN HUSSIN BIN WAN ISHAK**

Tandatangan
(Signature) :

Tarikh
(Date) : 17/10/06

PERMISSION OF USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from the 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 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 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 or in part, should be address to

Dean of Graduate School
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRACT

The performance of all technologies is highly depended on the quality of the data. For example, Neural Network (NN) technique can be applied very well if the data have been well prepared and free from noise and missing value. This study empirically compares several handling missing value methods for NN based on literature. Six of those methods have been identified and compared using adult data set (retrieved from UCI database). The methods include mean average, replace with one, replace with zero, replace with maximum, and replace with minimum and regression. The result shows that replace with maximum value method yield better accuracy compare to the other methods.

ACKNOWLEDGEMENT

Alhamdulillah....Praise to Allah for his guidance and blessing for giving me the strength and perseverance to complete this study.

I would foremost like to thank my parents, for providing me with the opportunity to pursue my goals and for their love and affection, which has helped me through the most trying times. Equal gratitude goes out to my siblings and my best friends.

Special thanks to my friends for their love, kindness, help and support throughout my stay in Malaysia.

I truly enjoyed the time we spent living together.
I would like to thank my supervisor:

Mr. Wan Hussain Bin Wan Ishak

For him guidance and constant motivation that has enabled me to complete my research work
I would also like to thank him for the opportunities that he has made available to me.

Lastly, I would like to thank Dr. Faudziah Ahmad for her guidance and support.

TABLE OF CONTENTS

List of Tables

List of Figures

REFERENCES

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1978). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion), *Journal of Royal Statistical Society*, vol.82, pp. 528-550.

Enders, C.K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8, pp. 128-141.

Enders, C.K, & Peugh, J.L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, 11, pp. 1-19.

Enqvist, A., Karlsson, G., Loviken, G., Möller, A. , Nilseng, AB, Nilsson, C. & Olsson, L. (2005). Methodology for handling forest industry environmental data – Method report, *Chalmers University of Technology.*

Fujikawa, Y. (2001). Efficient Algorithms for Dealing with Missing values in Knowledge Discovery. *Work paper series 01-11*, Re ts – o , sov Japan

Gan X., Liew A. W. & Yan1 H. (2006). Missing Microarray Data Estimation Based on Projection onto Convex Sets Method

Gold, M. S. & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, 7, pp. 319-355.

Graham, J.W. & Hofer, S.M. (2000). Multiple imputation in multivariate research. In T.D. Little, K.U. Schnabel, and J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches*, and specific examples (pp. 201-218).

Heitjan, D.F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4), pp. 548-550.

Howell, D.C. (2002). *Treatment of Missing Data*, retrieved from http://www.uvm.edu/~dhowell/StatPages/StatHomePage.html

Huisman, M. & Goudriaan, H. (2001). Handling missing item responses due to item nonresponse and incomplete designs. In J. Bethlehem & S. van Buuren (Eds.), pp. 57-73.

Information Technology Services at The University of Texas at Austin, http://Information Technology Services.htm, May 10, 2004.

Joseph L. S. & John W. G., (2002). Missing Data: Our View of the State of the Art, *Psychological Methods*, Inc. 2002, 7 (2), pp. 147–177.

Lakshminarayan K., Harp, S. A. and Samad, T. (1999). Imputation of Missing Data in Industrial Databases, *Applied Intelligence*, vol 11, pp. 259 – 275.

Little, R. J. A. & Rubin, D. A. (1987). Statistical analysis with missing data. John Wiley and Sons.

Little, R.J.A. & Rubin, D.B. (1989). The analysis of social science data with missing values, Sociological Methods and Research, 18, pp. 292-326.

Luo J, TaoYang & YanWang (2005). Missing Value Estimation For Microarray Data Based On Fuzzy C-means Clustering, School Computer and Communication, Hunan University, Changsha 410082, China pp. 611-616

Mitchell T.M. (1997). *Machine Learning*, McGraw-Hill,

Muhammad Shoaib B. S., Gondal I. & Dooley L. (2005). A Collateral Missing Value Estimation Algorithm For Dna Microarrays, Gscit, Monash University, VIC 3842, Australia IEEE

Muhammad Shoaib B. S., Gondal I. & Dooley L. (2005). K-Ranked Covariance Based Missing Values Estimation for Microarray Data Classification, Monash University, VIC 3842, Australia IEEE

Muthen, B., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 51, pp. 431-462.

Muthén, L.K. & Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 4, pp. 599-620.

Rahm E. & Do H. H. (2000). Data Cleaning: Problems and Current Approaches, *Bulletin of the IEEE Technical Committee on Data Engineering*, pp. 313.

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537-560.

Utsunomiya K. & Sonoda K. (2002). Methodology for Handling Missing Values in TANKAN, Research and Statistics Department, Bank of Japan, C.P.O BOX203 TOKYO, 100-8630 JAPAN

Vach, W. (1994), Missing Values: Statistical Theory and Computational Practice, In: P. Dirschedl, and R. Ostermann,(Eds.), *Computational Statistics*, Physica-Verlag, pp. 345-354.

Wagstaff K. L. & Laidler V. G. (2005). Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy, Astronomical Data Analysis Software and Systems XIV P2.1.25 , ASP Conference Series, V XXX, 2005

Wang, X.., Ao, L., Jiang, Z. & Feng, H. (2005). Novel method for missing value estimation in gene expression profile based on support vector regression, Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China

Yuan, K. H. & Bentler, P.M.. (2000). Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Non-Normal Missing Data, *Sociological Methodology*, pp.165-200.

Zhang, S., Qin, Z., Ling, C. & Sheng, S. (2005). Missing is Useful: Missing Values in Cost-sensitive Decision Trees, *IEEE Transactions on Knowledge and Data Engineering*, 17 (12), pp. 1689-1693.