COMPARING THE PERFORMANCES OF NEURAL NETWORK AND ROUGH SET THEORY TO REFLECT THE IMPROVEMENT OF PROGNOSTIC IN MEDICAL DATA

A thesis submitted to College of Arts and Sciences in partial fulfillment of the requirement for the degree Master of Science (Intelligent System) Universiti Utara Malaysia

By

Nur Aniza Bt Alang Ismail

December, 2009

COMPARING THE PERFORMANCES OF NEURAL NETWORK AND ROUGH SET THEORY TO REFLECT THE IMPROVEMENT OF PROGNOSTIC IN MEDICAL DATA

A thesis submitted to College of Arts and Sciences in partial fulfillment of the requirement for the degree Master of Science (Intelligent System) Universiti Utara Malaysia

By

Nur Aniza Bt Alang Ismail

December, 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 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.

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

Melalui penyelidikan singkat yang telah saya jalankan, saya telah menyelidik dua daripada teknik yang telah diperkerkenalkan dalam Kepintaran Buatan; iaitu Rangkaian Neural (Neural Network) dan juga Teori Set Kasar (Rough Set Theory). Kedua-dua teknik ini adalah dua teknik yang terbaik digunakan dalam penganalisaan data. Kepintaran Buatan adalah merupakan satu teknik yang masih deperingkat awal dan ianya baru diperkenalkan. Ianya masih lagi diperingkat pembangunan dan kegunaannya adalah menghasilkan sistem pintar yang dapat membantu manusia dalam kehidupan seharian bagi menyokong proses dalam membuat satu-satu keputusan.

Di Malaysia, Kepintaran Buatan adalah satu bidang yg masih lagi baru diperkenalkan. Satu kumpulan penyelidik dari Universiti Sains Malaysia telah menjalankan kajian tentang Kepintaran Buatan ini dalam bidang perubatan. Mereka juga bersetuju dengan kenyataan yang diberikan oleh para penyelidik Kepintaran Buatan seluruh negara bahawa Kepintaran Buatan sangat membantu dalam menggantikan kepintaran manusia. Dengan adanya elemen Kepintaran Buatan, ia membantu menyelesaikan pelbagai tugas manusia terutamanya dalam bidang perubatan dan disamping itu juga dapat mempercepatkan proses kerja seharian.

Dalam penyelidikan saya ini, saya telah memilih tiga set data perubatan iaitu; Ramalan Kanser Payudara dari Wisconsin, Penyakit Parkinson dan ramaln penyakit Hepatitis. Data-data perubatan ini telah dipilih kerana data-data yang berkaitan dengan perubatan seringkali digunakan oleh penyelidik-penyelidik Kepintaran Buatan dalam menjalankan kajian mereka. Selain itu, keputusan ramalan dan juga data-data yang digunakan dalam kajian ini mudah difahami. Selain itu, metodologi yang digunakan untuk kajian ini turut dibincangkan dan saya juga telah membuat kesimpulan dan juga kajian yang bakal dijalankan sebagai satu kesinambungan daripada kajian yang telah dijalankan ini.

ABSTRACT

In this research, I investigate and compared two of Artificial Intelligence (AI) techniques which are; Neural network and Rough set will be the best technique to be use in analyzing data. Recently, AI is one of the techniques which still in development process that produced few of intelligent systems that helped human to support their daily life such as decision making. In Malaysia, it is newly introduced by a group of researchers from University Science Malaysia. They agreed with others world-wide researchers that AI is very helpful to replaced human intelligence and do many works that can be done by human especially in medical area.

In this research, I have chosen three sets of medical data; Wisoncin Prognostic Breast cancer, Parkinson's diseases and Hepatitis Prognostic. The reason why the medical data is selected for this research because of the popularity among the researchers that done their research in AI by using medical data and the prediction or target attributes is clearly understandable. The results and findings also discussed in this paper. How the experiment has been done; the steps involved also discussed in this paper. I also conclude this paper with conclusion and future work.

ACKNOWLEDGEMENT

Alhamdulillah, it is with Allah S.W.T will that I get finish this Final Project in order to complete my Master's degree. I am very thankful to Dr. Fauziah bt Ahmad whom has been supervised me throughout this semester to complete this Master's Thesis. Also special thanks to Miss Aniza bt Mohamed Din whom helped me a lot in giving guidance and information in performing this thesis. Not forgetting my family for their support and understanding.

This paper is focusing in Artificial Intelligence techniques; Neural Network and Rough Set technique in order to get the best technique to be use in analyzing data. The software that has been used in the experiment is Neural Connection and ROSETTA.

"*To my beloved family, thanks for your support and sacrifice. To all my friends, nice knowing you all and thanks for the understanding and encouragement*."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
KDD	Knowledge Data and Discovery
NN	Neural Network
RS	Rough Set
OLAP	Online Analytical processing
HBV	Hepatitis virus B
WHO	World Health Organization
HBsAg	Hepatitis B surface Antigen
PNN	Probabilistic Neural Network
FNA	Fine needle aspirate
DM	Data Mining
MLP	Multilayer Perceptron
RMS	Root Mean Squared
GA	Genetic Algorithm

CHAPTER ONE: INTRODUCTION

Artificial Intelligence is one of approach that can train computers to think like human, where it can learn through experience, recognize patterns from large amount of data and also decision making process based from human knowledge and reasoning skills. According from an AI text book titled AI: Structures and Strategies for Complex Problem Solving, an AI can be defined as the branch of computer science that is concerned with the automation of intelligent behavior (Luger, 2005). It is combination of science and engineering field in order to make an intelligent machines, especially intelligent computer programs. There are three (3) perspectives in AI; 1) AI can be as a replacement, 2) as an assistant and 3) it also can be used to extend human capabilities (McCarthy J., 2007).

Nowadays, computers technology and data bases helps human in collecting and storing huge amount of data. The large size of most data bases makes it impossible for human to interpret data. Therefore, computers are needed for extracting new, useful knowledge. Lately, other science methods like machine learning, artificial intelligence and logics have made progress and achievements in this field. Today, as we can see the usage of Data Mining and Knowledge Discovery gives more advantages to Statisticians in order to reduce the information stored, to reduce costs, increase sales and revenues, also reduce accidents and failure within data (Dingsoyr T., 1997). There too many definitions about Data Mining and Knowledge Discovery

The contents of the thesis is for internal user only

REFERENCES

Cawsey, A. (1998). *The essence of artificial intelligence.* Prentice-Hall: England.

Depolo J. (2005). *Breath Easy.* Futures magazine. Vol. 23, No. 2. Retrieved from http://www.worldandijournal.com.eserv.uum.edu.my/subscribers/searchdetail.asp?num=25372

Legg S. and Hutter M. (2007). *Universal Intelligence: A Definition of Machine Intelligence*. Mind and Machines. Volume 17 , Issue 4. Kluwer Academic Publishers, Massachusetts.

Lehner P.E. (1986). *On the Role of Artificial Intelligence in Command and Control.* IEEE Transactions on Systems, Man, and Cybernetics, Vol. Smc-16, No. 6.

Luger, G.F (2005). *Artificial intelligence: structures and strategies for complex problem solving.* Fifth Edition. Addison-Wesley Publishing Corporation: New York.

McCarthy J. (2007). *What Is Artificial Intelligence?* Retrieved from http://www-formal.stanford.edu/jmc/

Reddy, R. (1996). *The Challenge of Artificial Intelligence. Computer, pp: 86-98.*

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

Lixiang Shen, Francis E. H. Tay, Liangsheng Qu and Yudi Shen (2000), *Fault Diagnosis using Rough Sets Theory* , Computers in Industry, vol. 43, Issue 1, 1 August 2000, pp.61-72.,

URL:www.geocities.com/roughset/Fault_diagnosis_using_rough_sets_theory.pdf

Israel E. Chen-Jimenez, Andrew Kornecki, Janusz Zalewski, *Software Safety*

Analysis Using Rough Sets,

URL:http://www-ece.engr.ucf.edu/~jza/classes/6885/rough.ps

Francis E. H. Tay and Lixiang Shen (2002), *Economic and Financial Prediction using Rough Sets Model,*, European Journal of Operational Research 141, pp.643-661, URL:http://www.geocities.com/roughset/EJOR.pdf

Pawan Lingras (2001), *Unsupervised Rough Set Classification Using GAs* Journal of Intelligent Information Systems, 16, 215–228, found on: CiteSeer, URL:http://citeseer.nj.nec.com/cs

Rapp, S., Jessen, M. and Dogil, G. (1994). *Using Rough Sets Theory to Predict German Word Stress*. in: Nebel, B. and Dreschler-Fischer, L. (Eds.) KI-94: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence 861, Springer-Verlag, URL:www.ims.uni-stuttgart.de/~rapp/ki94full.ps

Frasconi P. Soda G. and Vullo A.(2001). "*Text Categorization for Multiple page Documents: A Hybrid Naïve-Bayes HMM Approach"*. ACM-IEEE Joint Conference on Digital Libraries, 2001.

Chouchoulas A. and Shen Q (1999). "*A Rough Set-Based Approach to Text Classification"*. Proceedings of the 7th International Workshop on Rough Sets (Lecture Notes in Artificial Intelligence, No. 1711), pages 118-127, 1999.

Cunningham, Padraig (2004). "*Dimension Reduction and Feature Subset Selection*". Presentations MUSCLE Scientific Meeting Malaga, 4-5 Nov 2004.

Breault, Joseph L. (2001), "*Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?*" Computing Science and Statistics, 33, /I2001Proceedings/JBreault/JBreault.pdf

Bill B.W., McKay R.I., Abbas A.H. and Barlow M. (2000). "*A Comparative Study for Domain Ontology Guided feature Extraction"*. Proceedings of the twenty-sixth Australasian computer science conference on Conference in research and practice in information technology, p.69-78, February 01, 2003, Adelaide, Australia

Garg A. and Roth D. (2001). *"Understanding Probabilistic Classifiers"*. Conference Proceeding. ECML'01, Sept. 2001.

Jiye Li and Nick Cercone (2005). *"Empirical Analysis on the Geriatric Care Data Set Using Rough Set Theory"*. Technical Report, CS-2005-05, School Of Computer Science, University of Waterloo.

Marasek K. (1997). *"Methods of Data Classification"*. Experimental Phonetic Group. Institute of Natural Language Processing University of Stuttgart, Germany.

URL: http://www.ims.uni-stuttgart.dc/phonetic/EGG/pagev4.htm

Wang G.Y., Chen L. and Wu Y. (1999). "*Rough set based solutions for network Security*". IEEE Symposium on Security and privacy. IEEE Computer Society,1999.

Voges, K. E. (2005). "*Research techniques derived from rough sets theory: Rough classification and rough clustering*". Paper presented at ECRM2005: Fourth European Conference on Research Methods in Business and Management, April 21 - 22, 2005.

Toma I. (2004). "*Optimization Techniques for Neural Networks Text Classifiers*". Next Web Generation Seminar.Summer 2004.

Swiniarski R. W. (2001). "*Rough sets Methods in Feature Reduction and Classification*". International Journal Application Math. Computer Science, Vol. 11, No3, 565-582.

Strackeljan J. (1999) "*Feature Selection Methods for Soft Computing Classification*". Proceedings of the ESIT 1999 (European Symposium on Intelligent Techniques), Kreta, Juni 1999.

Stone T. (2003). "*Parameterization of naïve Bayes Spam Filtering*". Masters Comprehensive Exam, University of Colorado at Boulder, 2003.

Steppe J.M. (1994). "*Feature and Model Selection in Feed Forward Neural Networks*". PhD Dissertation at Air Force Institute of Technology, Ohio.

Provost J. (2002). "*Naïve-Bayes vs Rule-Learning in Classification of Email*". The University of Texas at Austin, Artificial Intelligent Lab. Technical report AITR-99-284.

Pearl Judea (1988). *"Probabilistic Reasoning in Intelligent Systems"*. San Mateo, CA:Morgan Kaufman Publishers.

Pechenizkiy M, Puuronen S. and Tsymbal A. (2000). "*Feature Extraction for Classification in Knowledge Discovery Systems*". In: V.Palade, R.J.Howlett, L.C.Jain (Eds.), Proc. 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems KES'2003, Lecture Notes in Artificial Intelligence, Vol.2773, Heidelberg: Springer-Verlag, pp. 526-532.

Painter James (2003). *"Uses of Bayesian Statistics"*. Referred on 12th October 2005 URL: http://www.tessella.com/Literature/Supplements/PDF/bayesianstats.pdf.

Pawlak Zdzislaw (1982). "*Rough Sets*". International Journal of Computer and Information Sciences 11 (1982): 341-356.

Pearl Judea (1988). *"Probabilistic Reasoning in Intelligent Systems"*. San Mateo, CA: Morgan Kaufman Publishers.

Orhn, A. and J. Komorowski (1997). "*ROSETTA: A Roughset Toolkit for Analysis of Data*". Joint Conference of Information Sciences: semiotics, fuzzy logic, soft computing, computer vision, neural computing, genetic algorithm, pattern recognition, evolutionary computing, Durham NC, Duke University Press: 403-407.

Lewis D.D. and Ringuette M. (1995). *"Comparison of Two Learning Algorithms for text Categorization"*. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR' 94).

Grenager H. S. (1996). "*Rough Sets*". Referred on 19th P October 2005.

URL: http://www.pvv.ntnu.no/~hgs/project/report/node38.html

Yao Y.Y., (1998). *"A comparative study of fuzzy sets and rough sets"*. Journal of Information Sciences, vol. 109, pp. 227-242.

Slowinski, K., Slowinski, R., Stefanowski, J., (1988). *"Rough set approach to analysis of data from peritoneal lavage in acute pancreatitis"*. Medical Informatics 13, pp. 143-159.

Shusaku Tsumoto. (1998). *"Knowledge discovery in medical databases based on rough sets and feature-oriented generalization"*. IEEE World Congress on Fuzzy Systems Proceedings, Vol. 2, Issue , 4-9 May 1998. pp.1296 - 1301.

Roman W. Swiniarski, *"Rough Sets and Bayesian Methods Applied to Cancer Detection"*. Rough Sets and Current Trends in Computing: First International Conference, RSCTC'98, Warsaw, Poland, June 1998, pp.609-616.

M. Brameier, W. Banzhaf, *"A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining"*. IEEE Transactions on Evolutionary Computation, 2000.

Aleksander Øhrn. *"ROSETTA Technical Reference Manual"*. Knowledge Systems.

I. Dűntsch. G. Gediga (1999). *"Rough set data analysis: A road to non-invasive knowledge discovery"*.

Lin, T.Y., and N. Cercone (1997). "*Rough Sets and Data Mining"*. Kluwer Academic Publishers.

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

Gunther, G., and D. Ivo (2000). "*Statistical Techniques for Rough Set Data Analysis in Rough Sets: New Developments"*. Physica–Verlag.

Lavrajc, N., E. Keravnou and B. Zupan (1997). "*Intelligent Data Analysis in Medicine and Pharmacology"*. Kluwer Academic Publishers.

Zhong, N., and A. Skowron (2000). *"Rough sets in KDD: tutorial notes"*. Bulletin of International Rough Set Society, 4(1/2).

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

Ian H. Witten and Eibe Frank. *"Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)"*. Morgan Kaufmann Publishers Inc., 2005.

Zdzislaw Pawlak. *"Rough Sets: Theoretical Aspects of Reasoning about Data"*. Kluwer Academic Publishers, Norwell, MA, USA, 1992.