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

Copyright © Nur Aniza Bt Alang Ismail, 2009
All rights reserved
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 diperkenalkan 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 sehari-hari bagi menyokong proses dalam membuat satu-satu keputusan.

Di Malaysia, Kepintaran Buatan adalah satu bidang yang 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 ke pintaran manusia. Dengan adanya elemen Kepintaran Buatan, ia membantu menyelesaikan pelbagai tugas manusia terutamanya dalam bidang perubatan dan disamping itu juga dapat mempercepatkan proses kerja sehari-hari.

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; Wisconcin 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

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERMISSION TO USE</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td></td>
<td>II</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td></td>
<td>III</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td></td>
<td>IV</td>
</tr>
<tr>
<td>DEDICATION</td>
<td></td>
<td>V</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td></td>
<td>VI</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>XIII</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td></td>
<td>XIV</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Problem Background and Problem Statement</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>Research Objectives</td>
<td>8</td>
</tr>
<tr>
<td>1.3</td>
<td>Scope</td>
<td>9</td>
</tr>
<tr>
<td>1.4</td>
<td>Project Significance and Contributions</td>
<td>10</td>
</tr>
<tr>
<td>1.5</td>
<td>Conclusions</td>
<td>10</td>
</tr>
</tbody>
</table>
2 LITERATURE REVIEW

2.0 Literature Review 12
2.1 Knowledge Discovery 12
2.2 Neural Network 14
2.3 Rough Set 17
2.4 Hepatitis 18
2.5 Parkinson 20
2.6 Breast Cancer 21
2.7 Conclusions 22

3 METHODOLOGY

3.0 Methodology 24
3.1 KDD Stages 25
   3.1.1 Selection 25
   3.1.2 Data Cleansing and Pre-Processing 25
   3.1.3 Data Mining 26
   3.1.4 Interpretation and Evaluation 26
3.2 Proposed Methodology 27
3.3 Network Network 27
   3.3.1 System Analysis 27
   3.3.2 Pre-Processing 37
3.3.3 Data Mining Algorithm/Technique  40
3.3.4 Post-Processing  44

3.4 Rough Set technique
3.4.1 Pre-Processing  47
3.4.2 Data Mining Algorithm/Technique  48
3.4.3 Post-Processing  49
3.4.4 Testing Process  51

3.5 Conclusions  52

4  EXPERIMENT AND FINDINGS

4.0 Experiment and Findings  53
4.1 Neural Connection  53
4.1.1 The Experiment: Wisconsin Prognostic Breast Cancer  53
4.1.2 The Experiment: Parkinson’s Diseases  68
4.1.3 The Experiment: Hepatitis Diseases  80
4.2 ROSETTA Toolkit  93
4.2.1 The Experiment: Wisconsin Prognostic Breast Cancer  93
4.2.2 The Experiment: Parkinson’s Diseases  98
4.2.3 The Experiment: Hepatitis Diseases  103
4.3 Conclusions 108

5 DISCUSSION OF RESULTS

5.0 Discussion 109

5.1 Neural Network 109

5.2 Rough Set 111

5.3 Conclusions 112

6 CONCLUSIONS

6.0 Conclusions 113

6.1 Project Contributions 113

6.2 Project Advantage 114

6.3 Suggestions and Future Works 114

REFERENCES 115
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Data information about Wisconsin Prognostic Breast Cancer</td>
<td>28</td>
</tr>
<tr>
<td>3.2</td>
<td>Data information about Parkinson’s diseases</td>
<td>32</td>
</tr>
<tr>
<td>3.3</td>
<td>Data information about Hepatitis diseases</td>
<td>35</td>
</tr>
<tr>
<td>3.4</td>
<td>Descriptive Statistics for Wisconsin Breast Cancer</td>
<td>38</td>
</tr>
<tr>
<td>3.5</td>
<td>Descriptive Statistics for Hepatitis diseases</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Result to determine the best hidden unit</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>Result to determine the best hidden unit</td>
<td>57</td>
</tr>
<tr>
<td>4.3</td>
<td>Result to determine the most suitable learning rate</td>
<td>59</td>
</tr>
<tr>
<td>4.4</td>
<td>Result to determine the best learning rate</td>
<td>60</td>
</tr>
<tr>
<td>4.5</td>
<td>Result to determine the suitable momentum rate</td>
<td>61</td>
</tr>
<tr>
<td>4.6</td>
<td>Result to determine the best momentum rate</td>
<td>62</td>
</tr>
<tr>
<td>4.7</td>
<td>Result to determine suitable activation function</td>
<td>63</td>
</tr>
<tr>
<td>4.8</td>
<td>Result to determine the suitable stopping criteria or epoch</td>
<td>65</td>
</tr>
<tr>
<td>4.9</td>
<td>Result to determine the best stopping criteria</td>
<td>66</td>
</tr>
<tr>
<td>4.10</td>
<td>Result to determine the suitable hidden unit</td>
<td>70</td>
</tr>
<tr>
<td>4.11</td>
<td>Result to determine the best hidden unit</td>
<td>71</td>
</tr>
<tr>
<td>4.12</td>
<td>Result to determine the most suitable learning rate</td>
<td>73</td>
</tr>
</tbody>
</table>
4.13 Result to determine the best learning rate
4.14 Result to determine the best activation function
4.15 Result to determine the suitable number of epoch
4.16 Result to determine the best suitable number of epoch
4.17 Result to determine the suitable hidden unit
4.18 Result to determine the best hidden unit
4.19 Result to determine the suitable learning rate
4.20 Result to determine the best learning rate
4.21 Result to determine the suitable momentum rate
4.22 Result to determine the best momentum rate
4.23 Result to determine the activation function
4.24 Result to determine the suitable number of epoch
4.25 Result to determine the best number of epoch
4.26 Results using Boolean reasoning algorithm
4.27 Results using Entropy/MDL algorithm
4.28 Results using Equal Binning frequency
4.29 Results using Naïve algorithm
4.30 Results using Semi Naïve algorithm
4.31 Results using Boolean reasoning algorithm
4.32 Results using Entropy/MDL algorithm
4.33 Results using Equal Frequency binning 100
4.34 Results using Naïve algorithm 101
4.35 Results using Semi Naïve algorithm 102
4.36 Results using Boolean reasoning algorithm 103
4.37 Results using Entropy/MDL algorithm 104
4.38 Results using Equal Frequency binning 105
4.39 Results using Naïve algorithm 106
4.40 Results using Semi naïve algorithm 107
4.41 Comparisons of accuracy between Neural network and Rough set techniques 108
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURES</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Artificial Neural Network</td>
<td>3</td>
</tr>
<tr>
<td>3.1</td>
<td>An overview of the steps in KDD process</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Methodology for Medical Prognostic Using Data Mining</td>
<td>27</td>
</tr>
<tr>
<td>3.3</td>
<td>Sample of Breast Cancer data from Neural Connection tool kit</td>
<td>30</td>
</tr>
<tr>
<td>3.4</td>
<td>Pie chart for class distribution (Wisconsin Prognostic Breast Cancer)</td>
<td>30</td>
</tr>
<tr>
<td>3.5</td>
<td>Pie chart for class distribution (Parkinson’s diseases)</td>
<td>34</td>
</tr>
<tr>
<td>3.6</td>
<td>Pie chart for class distribution (Hepatitis diseases)</td>
<td>36</td>
</tr>
<tr>
<td>3.7</td>
<td>Process flow diagram for Multilayer Perceptron</td>
<td>40</td>
</tr>
<tr>
<td>3.8</td>
<td>Sequential data for Wisconsin Breast Cancer</td>
<td>41</td>
</tr>
<tr>
<td>3.9</td>
<td>Data allocation in Neural Connection toolkit</td>
<td>42</td>
</tr>
<tr>
<td>3.10</td>
<td>Parameters control for Neural Network</td>
<td>43</td>
</tr>
<tr>
<td>3.11</td>
<td>Other parameters setting for backpropagation learning</td>
<td>44</td>
</tr>
<tr>
<td>3.12</td>
<td>Data that has been discretized</td>
<td>47</td>
</tr>
<tr>
<td>3.13</td>
<td>Result after discretization process</td>
<td>50</td>
</tr>
<tr>
<td>3.14</td>
<td>Result of reduction process where rules being generated</td>
<td>51</td>
</tr>
<tr>
<td>3.15</td>
<td>Result of accuracy measurement</td>
<td>52</td>
</tr>
</tbody>
</table>
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


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


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


