MINING STUDENTS’ DATA WITH HOLLAND MODEL
USING NEURAL NETWORK AND LOGISTIC
REGRESSION

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

by

Noorlin binti Mohd Ali

© Noorlin binti Mohd Ali, 2005. All rights reserved.
NOORLIN BINTI MOHD. ALI

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)

MINING STUDENTS’ DATA WITH HOLLAND MODEL USING NEURAL NETWORK AND LOGISTIC 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
and that the project paper acceptable in form and content, and that a satisfactory
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) : [Signature]
Tarikh (Date): [Signature]

Nama Penyelia Kedua
(Name of 2nd Supervisor): MISS NOORAINI YUSOFF

Tandatangan
(Signature) : [Signature]
Tarikh (Date): [Signature]
PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for the postgraduate degree from Universiti Utara Malaysia, I agree that 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 or, in their absence by the Dean of 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 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 addressed to:

Dean of Faculty of Information Technology
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman
ABSTRAK (BAHASA MELAYU)

Bidang pendidikan mempunyai banyak aplikasi perlombongan data yang menarik dan mencabar, serta dikenalpasti sebagai satu alat yang berpotensi digunakan untuk membantu tenaga pengajar dan pelajar, dan memperbaiki kualiti sistem pendidikan. Kesan pengumuman Menteri Pendidikan Tinggi mengenai lebihan graduan terutamanya dari universiti awam secara tidak langsung turut memberi kesan kepada pengambilan/kemasukan pelajar ijazah sarjana muda di Universiti Utara Malaysia (UUM). Sehubungan itu, pelajar yang mengikut program di Fakulti Teknologi Maklumat (FTM) dan Fakulti Pengurusan Teknologi (FTP) mempunyai pelbagai latarbelakang pendidikan. Justeru, kajian ini bertujuan untuk meninjau latarbelakang pelajar tahun pertama yang mengambil program Ijazah Sarjana Muda Teknologi Maklumat (Bachelor of Information Technology-BIT), Ijazah Sarjana Muda Multimedia (Bachelor of Multimedia-BMM), dan Ijazah Sarjana Muda Pengurusan Teknologi (Bachelor of Management of Technology-BMoT) di UUM. Di samping itu, Model Personaliti Holland turut diaplikasikan bagi mengenalpasti jenis personaliti pelajar. Hasil kajian mendapati pelajar BIT bukan dari kumpulan Social kerana tiada nilai signifikan ke atas salan-soalan dari kumpulan Social. Kebanyakan pelajar BIT merupakan pelajar dari latarbelakang Sastera kecuali beberapa orang pelajar yang pernah mengambil dan menduduki subjek Perkomputeran di peringkat Sijil Tinggi Pelajaran Malaysia (STPM). Dari sudut Model Holland pula, pelajar BIT dirumuskan sebagai Artistic, Investigative, Realistic (ARI). Pelajar didapati lebih bersifat Artistic berdasarkan 50% daripada soalan-soalan yang diberikan untuk mengenalpasti personaliti pelajar adalah signifikan. Di samping itu, pelajar juga didapati terdiri daripada kumpulan Investigative (33.33%) dan Realistic (33.33%). Hasil kajian ini adalah selari dengan teori Holland berdasarkan kajian Hansen dan Campbell (1985) yang merumuskan kod personaliti bagi bidang komputer ialah Investigative, Realistic, dan Artistic (IRA).
Education domain provides many interesting and challenging in data mining applications that potentially identified as a tool to help both educators and students, and improve the quality of education system. Nowadays, the impact of Minister of Education (MOE) regarding surplus graduates particularly from public universities somehow had an impact on Universiti Utara Malaysia’s (UUM) undergraduate intake. As a result, students who applied to undertake a program at Faculty of Information Technology and Faculty of Management Technology come from various background. Hence this study aims to get some insight into first year students undertaking undergraduate program such as Bachelor of Information Technology (BIT), Bachelor of Multimedia (BMM) and Bachelor in Management of Technology (BMOT) at Universiti Utara Malaysia. The Holland Personality Model was used to indicate the students' personality traits. The study concluded that BIT students are not from the Social type since none of the Social personality type is significant. Most of BIT students have Arts background, except a few who have sat for Perkom (Perkomputeran) subject during the STPM examination. As for the Holland Model, It also appears that BIT students are more Artistic since 50% of the questions that measure the personality type is significant. In addition, the BIT students are Realistic (33.33%) and Investigative (33.33%) type. The results also reveal that the BIT students concluded as Artistic, Investigative and Realistic (AIR) in personality types that are in accordance to Holland personality theory, this finding were also supported by Hansen and Campbell (1985) that suggested that Investigative, Realistic and Artistic (IRA) should be the code for computer professionals.
ACKNOWLEDGEMENTS

In the name of Allah, Most Gracious, Most Merciful. Peace upon the prophet, Muhammad S.A.W. Alhamdulillah, a foremost praise and thankful to Allah for His blessing, giving me the strength in completing this study.

My endless appreciation goes to both of my respective supervisors; Associate Professor Fadzilah Siraj and Miss Nooraini Yusoff for the guidance, patience, encouragement, advice and flourish of knowledge during completing these three semesters course.

My warm appreciation dedicates to the lecturers of Department of Computer Science UUM, the student of MSc. Intelligent Systems (June 2004 and November 2003 batches) and all of my friends for all of the knowledge, advice and moment we've shared. My special thanks also goes to Haji Aris Zainal Abidin, Rahmatul Hidayah Salimin, Kak Aai, Kak Lily.

The first, last and always, a lasting heartfelt gratitude to my mother, Inah binti Haji Hassan for all of the love, du'a and support in completing this course, as well as to Long, Ngah, Diya and J.

Special thanks to the respondents and lecturers for the cooperation during data collecting session for this study.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>DESCRIPTIONS</th>
<th>PAGE NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERMISSION OF USE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRAK (BAHASA MELAYU)</td>
<td>ii</td>
</tr>
<tr>
<td>ABSTRACT (ENGLISH)</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>x</td>
</tr>
</tbody>
</table>

## CHAPTER ONE: INTRODUCTION

1.1 Background 1
1.2 Problem Statement 6
1.3 Project Objectives 7
1.4 Significance of the Study 7
1.5 Project Scope 8
1.6 Thesis Organization 8

## CHAPTER TWO: LITERATURE REVIEW

2.1 Data Mining 9
2.2 Neural Networks 10
2.3 Regression Analysis 13
2.4 Applications of NNs and Statistical in forecasting 15
   2.4.1 Neural Networks in Education 17
   2.4.2 Statistical Analysis in Education 21
2.5 Personality Psychology 24
   2.5.1 Holland Hexagonal Personality Model 28
2.6 Summary 31
CHAPTER THREE: NEURAL NETWORK, HOLLAND PERSONALITY MODEL AND METHODOLOGY

3.1 Networks Architecture
3.2 Training Method
3.2.1 Supervised Learning
3.2.2 Unsupervised Learning
3.3 Backpropagation Algorithm
3.3.1 Backpropagation Architecture and Algorithm
3.3.2 Learning Parameter
  □ *Learning Rate*
  □ *Momentum Rate*
3.4 Building Neural Networks Forecasting Model
3.5 Holland Hexagonal Personality Model
  3.5.1 Categorizations of Holland Personality Theory
    □ *Realistic (R)*
    □ *Investigative (I)*
    □ *Artistic (A)*
    □ *Social (S)*
    □ *Enterprising (E)*
    □ *Conventional (C)*
3.6 Methodology
  3.6.1 Instrumentation
  3.6.2 Variable Selection
  3.6.3 Data Collection
    □ *Data Acquisition*
    □ *Data Description*
  3.6.4 Data Preprocessing
    □ *Data Cleaning*
    □ *Data Transformation*
    □ *Output Representation*
  3.6.5 Training, Testing and Validation Sets
  3.6.6 Neural Network Paradigm
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6.7</td>
<td>Evaluation Criteria</td>
<td>65</td>
</tr>
<tr>
<td>3.6.8</td>
<td>Regression Model of Student’s Data</td>
<td>65</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary</td>
<td>66</td>
</tr>
</tbody>
</table>

### CHAPTER FOUR: RESULTS AND FINDINGS

4.1 The Convenient Sampling Dataset 67

4.2 The Experiments on STPM's results subjects 69

4.3 The Experiments on Holland Model 74

### CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Conclusion 78

5.2 Problems and Limitations 80

5.3 Recommendation 81

### REFERENCES

82

### APPENDIXES

Appendix A: Sample of raw data 90

Appendix B: Sample of Questionnaire 98
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 3.1</td>
<td>A single layer networks architecture</td>
<td>34</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Multilayer networks architecture</td>
<td>34</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>A recurrent networks architecture</td>
<td>35</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>A backpropagation network with three layers</td>
<td>38</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>The diagram of backpropagation neural network for modeling student program based on STPM’s result and Holland personality test</td>
<td>45</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>The summarization of Holland’s six personality types</td>
<td>47</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>The Steps in Performing Neural Network Experiments</td>
<td>56</td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>The neural network structure for modeling student program based on STPM’s result and Holland personality test</td>
<td>64</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>The percentage distribution of respondents based on the program</td>
<td>68</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>The mean value of STPM examination for each subject</td>
<td>69</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>The mean value for STPM subject after combination</td>
<td>70</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>The percentage of before and after combining subject</td>
<td>71</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>Mean value for STPM students based on the BMM, BMoT and BIT program</td>
<td>72</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1</td>
<td>The questions on Artistic type</td>
<td>54</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>The questions on Realistic type</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>The questions on Social type</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>The questions on Investigative type</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>The questions on Enterprising type</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>The questions on Conventional type</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>The list of grade point value for STPM examination</td>
<td>59</td>
</tr>
<tr>
<td>Table 3.8</td>
<td>The value representation for each answer in Holland personality test</td>
<td>60</td>
</tr>
<tr>
<td>Table 3.9</td>
<td>Sample of students' datasets before the normalization</td>
<td>61</td>
</tr>
<tr>
<td>Table 3.10</td>
<td>Sample of students' datasets after the normalization</td>
<td>61</td>
</tr>
<tr>
<td>Table 3.11</td>
<td>Output Representation</td>
<td>61</td>
</tr>
<tr>
<td>Table 3.12</td>
<td>Data Distribution for Student Dataset</td>
<td>62</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>The Total number of respondents based on the selected undergraduate program</td>
<td>67</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>The comparison percentage of NN and Logistic Regression</td>
<td>70</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>The comparison of both method before and after combining subjects</td>
<td>71</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>The significant value of each subject</td>
<td>71</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>The result of NN and Logistic Regression with and without the combination of Perkomp subject</td>
<td>73</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>The significant value of each subjects</td>
<td>73</td>
</tr>
<tr>
<td>Table 4.7</td>
<td>The comparison of both method on Holland Model</td>
<td>74</td>
</tr>
<tr>
<td>Table 4.8</td>
<td>The comparison of both method with the combination of result and Holland Model</td>
<td>74</td>
</tr>
<tr>
<td>Table 4.9</td>
<td>NN Model obtained from students' data</td>
<td>75</td>
</tr>
<tr>
<td>Table 4.10</td>
<td>The result of Logistic Regression to the selected dataset</td>
<td>76</td>
</tr>
</tbody>
</table>
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>Data Mining</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>Multilayer Perceptron</td>
<td></td>
</tr>
<tr>
<td>STPM</td>
<td>Sijil Tinggi Pelajaran Malaysia</td>
<td></td>
</tr>
<tr>
<td>BIT</td>
<td>Bachelor of Information Technology</td>
<td></td>
</tr>
<tr>
<td>BMM</td>
<td>Bachelor of Multimedia</td>
<td></td>
</tr>
<tr>
<td>BMoT</td>
<td>Bachelor of Management of Technology</td>
<td></td>
</tr>
<tr>
<td>UUM</td>
<td>Universiti Utara Malaysia</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER ONE

INTRODUCTION

This section discusses the background of the study that consists of general overview on data mining techniques, which have been used in this study. A brief description on the selected domain, education domain is also reviewed. The section also consists of the problem statement, list of project objectives, significance of the study conducted, and the study scope. Finally, this section presents the thesis organization that describing the structure of this report.

1.1 Background

Data mining (DM) has been extensively investigated for potential applications in many domains. It is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematical algorithms, and statistics (Liao, 2003). The field of data mining and
The contents of the thesis is for internal user only
REFERENCES


NJ: Prentice Hall.


87


System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on


