ROUGH SET RULES EXTRACTION FOR STUDENT PROGRAMMING SKILLS

MOKHTAR MASSOUD KERWAD

UNIVERSITI UTARA MALAYSIA
2006

Rough Set Rules Extraction for Student programming skills

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

By

Mokhtar Massoud Kerwad

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)

MOKHTAR MASSOUD KERWAD

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)

ROUGH SET RULES EXTRACTION FOR STUDENT PRORAMMING SKILLS

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. AZIZI AB AZIZ**

Tandatangan
(Signature) : __________ Tarikh (Date): 15/08/06

Nama Penyelia Kedua
(Name of 2nd Supervisor): **MR. MOHD. SHAMRIE SAININ**

Tandatangan
(Signature) : __________ Tarikh (Date): 15/08/06

PREMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from the university 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 part, for scholarly purposes may be granted by my supervisor or, in their absence by the Dean of the Faculty of Information Technology. It is understand 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 understand that due recognition shall be given to me and to University 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 whol or in part, should be addressed to.

Dean of Faculty of Information technology
University Utara Malaysia
06010 Sintok
Kedah Darul Aman.

ABSTRACT

Programming is a critical subject to computer science or information technology students. It is one of the fundamental skills they need to acquire during study. The aim of the study is to generate a compact set of rules using real data to predict student's performance. Not all variables as usual if good results are to be obtained. Data mining refers to one of the phases or step within the knowledge discovery in databases (KDD) processes for extracting used rough set technique. The extracted rules will be a measurement of the students' performance in programming and give the insight to educators on what should be help the students to master programming skills.

ACKNOWLEDGMENT

By the Name of Allah, the Most Gracious and the Most Merciful

Above others, my praise to Allah S.W.T whose blessing and guidance have helped me through entire project work. Peace be upon our Prophet Mohammed S.A.W, how has given light to mankind. My highest and most sincere appreciation goes to my beloved parents (Massoud Kerwad and Mona Mokhtar) and other family members for their patience, prayers and understanding over entire period of my study, although I was away from home but their care and concerned never make me left alone and always gave me love and encourage me along the way.

I would also not to waste this precious opportunity to express my deep appreciation and deepest gratitude to my supervisors, Mr. Azizi Bin Ab Aziz, Faculty of Information Technology, Universiti Utara Malaysia (UUM) for his full support and assistant throughout this study. I wish to acknowledge his assistance and time, provided excellent facilities, support and guidance throughout the project also for his advice during this project. An appreciation also goes to Dr. Faudziah Ahmad, I have learned much about rough set software.

Last but not least, a very special thanks to all my beloved best friends whose names not need to be addressed here they are always in my heart. Despite their indirect involvement, the bonds of brothers they have built have created a convenient yet supportive environment for me to successfully complete the study. To thanks to all the lecturers and members of MSc. Thanks again to everyone including those who I have probably forgotten to mention here. Intelligent System batch July 2006, all the best.

Mokhtar Massoud Kerwad
Faculty of Information Technologh
Department of Computer Science
University Utara Malaysia
July 2006

TABLE OF CONTENT

Page

CHAPTER 1: INTRODUCTION

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: METHODOLOGY

CHAPTER 4: FINDINGS AND RESULTS

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Chapter 1

CHAPTER 1

INTRODUCTION

In the real world, organizations often have large amount of data that are stored in databases. The large size of data makes data analysis difficult, as data are more complex in terms of number of attributes and number of objects. The use of a sufficient number of attributes and objects are one way to overcome the problem. In data mining, there are many techniques that can be used for reducing data. Such as, Rough set, multi discriminate analysis (MDA), classification and regression tree (CART) and principal component analysis (PCA). However, most reduction techniques perform differently when applied to various problems. To date there is no research that can identify which reduction technique is the best. This is because one reduction technique may be suitable to be used on one problem domain but unsuitable when applied on another problem domain.

There have been very few studies in recent years into academic success in computer programming. Today, despite of job saturation, industry is still keen to accept IT graduates and its main focus will be on any bright students can that do programming well. But many students who are proficient in many non-programming fail to achieve success in programming (Byrne and Lyons, 2001).

REFERENCES

Beynon, M., Curry, B., and Morgan, P., (2000). Classification and rule induction using Rough set theory. Expert system 17(3) 136-148.

Byrne, P., and Lyons, G., (2001). The Effect of Student Attributes on Success in Programming. *Department of Information Technology, ACM ISBN.*

Bergin, S., and Reilly, R., (2005). Programming: Factors that Influence Success, *ACM SIGCSE*, pp. 411-415.

Bilski, P., Walczak, Z. and Wojeiechowski, J., (2005). Diagnostics of analog systems using Rough sets. Institute of Radioelectronics, Warsaw University of Technology, ul, Nowowiejska 15/19 Warsaw, Poland.

Dash, M., Liu, H. and Yao, J. (1997), Dimensionality Reduction of Unsupervised Data. IEEE, Department of Information Systems and Computer Science.

Dunham, M. H., (2003). *Data Mining: Introductory and Advanced Topic*. Upper Saddlę River, NJ: Prentice Hall.

Evans, G.E. and Simkin, M.G., (1989). What Best Predicts Computer Proficiency. *Communications of the ACM*, Volume 32 Number 11.

Facey-Shaw, L., and Golding, P., (2005). Effects of Peer Tutoring and Attitude on Academic Performance of First Year Introductory Programming Students. *ASEE/IEEE Frontiers in Education Conference S1E-2.*

Fayyad, U., Shapiro, G. P., and Smyth, P., (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the acm november 1996/vol. 39, no. 11.

Grossman, R., Kasif, S., Moore, R., Rocke, D., and Ullman, J., (1999). A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data. *http://www.ncdm.uic.edu/m3d2.htm.*

Hostetler, T., R., (1983). Predicting Student Success in an Introduction Programming Course. Proceedings of NECC5. *IEEE Pre*ss, Silver Spring, Md.

Hu, X., and Cercone, N., (1994). Discovery of Decision Rules in Relational Databases: A Rough set Approach, *ACM Department of Computer Science*.

Han, J. and Kamber, M., (2001). Data Mining Concepts and Techniques. Morgan Kaufmann. San Francisco.

Hu, X., Lin, T., Y., and Han, J., (2004). A New Rough sets Model Based on Database Systems, *Fundamenta Informaticae* XX (2004) 1–18.

Honghai, F., Guoshun,, C., Yufeng, W., Bingru, Y., and Yumei, C., (2005). Rough set Based Classification rules generation for SARS Patients. *IEEE.*

Jiang, Y., Xu, C., Gou., J and Li, Z., (2004). Research on Rough set Theory Extension and Rough Reasoning. *IEEE, International Conference on Systems.*

Kusiak, A., (2001). Rough set Theory: A Data Mining Tool for Semiconductor Manufacturing. *Department of Industrial Engineering, Intelligent Systems Laboratory.*

Lil, J., and Cercone, N., (2005). A Rough set Based Model to Rank the Importance of Association Rules Vol.3642, pp.109-118.

Ma, Y., Liu, B., Wong, C. K., Yu, P.s., and Lee, S. M., (2000). Targeting the Right Students Using Data Mining, *ACM,* pp. 457-463.

Miller, M. T., Jerebko, A. K., Malley, J. D., and Summers, R. M., (2003). Feature Selection for Computer- Aided Polyp Detection using Genetic Algorithms. Proceedings of SPIE Vol 5031.

Norita, M., N, Hibadullah, C., F., and Osman, J., (2005). Factors Affecting Performance in Introductory Programming. Faculty of Information Technology, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.

Pawlak, Z., (1996). Rough sets and Data Analysis. Institute of Theoretical and Applied Informatics Polish Academy of Sciences, *IEEE*

Pillay, N., Vikash R., and Jugoo., (2005). An Investigation into Student Characteristics Affecting Novice Programming Performance. *Department of Computer Science.* Volume 37, Number 4.

Peng, Y., Liu, G., Lin, T and Geng, H., (2005). Application of Rough set Theory in Network Fault Diagnosis. *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05) IEEE.*

Raymer, M. L., Punch,W. F., Goodman, E. D., Kuhn, L. A., and Jain, A. k., (2000). Dimensionality Reduction Using Genetic Algorithms. *IEEE Transactions on Evolutionary Computation,* vol. 4, no. 2.

Ronald D., and McFarland., (2003), Teaching Students to Learn in the Computer Science and Information Systems Curriculum: Creating a Distinction Between Content and Methods. JCSC 19,1 (October 2003).

Revett, K., Gorunescu, F., Gorunescu, M., Darzi, E., and Ene, M., (2005). A Breast Cancer Diagnosis System: A Combined Approach Using Rough sets and Probabilistic Neural Networks. *IEEE*

Schultz, M. G., Eskin, E., and Stolfo, S. J., (2000), Data Mining Methods for Detection of New Malicious Executables, Department of Computer Science, Colombia University and Stat University of New york.

Tan, P., Steinback, M. and Kumar, V., (2006). Introduction to Data Mining. Addison Wesley. Pearson Education.

Vaishnavi, V. and Kuechler, W., (2005). Design Research in Information Systems. Retrieved Fabruary 20, 2004, last updated June 5, 2005, from http://www.isworld.org/Researchdesign/drisISworld.htm.

Werth, L. H., (1986), Predicting Student Performance in a Beginning Computer Science Class. Department of Computer Sciences, *ACM.*

Wei, J., Huang, D., Wang, S., and Ma, Z., (2002), Rough set Based Decision tree. Intelligent Control and Automation, 2002. Proceedings of the 4th World Congress on. Volume 1, Page(s):426 - 431 vol.1 Digital Object Identifier.

Yusof, A. M., and Abdullah, R., (2005), The Evolution of Programming Courses: Course curriculum, students, and their performance. *Computer Science and Information Technology*. Volume 37, Number 4.