CLASSIFICATION AND REGRESSION TREE IN CLASSIFYING AND PREDICTING STUDENTS' ACADEMIC PERFORMANCE

HO SU JUIH

UNIVERSITI TEKNOLOGI MALAYSIA

CLASSIFICATION AND REGRESSION TREE IN CLASSIFYING AND PREDICTING STUDENTS' ACADEMIC PERFORMANCE

HO SU JUIH

A thesis submitted in fulfillment of the requirements for the award of the degree of Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

JANUARY 2013

DEDICATION

I would like to dedicate this thesis to my beloved family:

My dearest father: Ho Aik Pah

My dearest mother: Chai Chiew Hiong

My dearest brothers: Ho Howe Sin
Ho Howe Hong
Ho Howe Kent

A**CKNOWLEDGEMENT**

Firstly, I would like to express my thankfulness to my supervisor, Dr. Norhaiza Bt. Ahmad for guiding me through the whole thesis. She has provided many useful materials and helped with implementation problems I encountered during the development. She also helped me revise over the thesis draft. Moreover, she gave me valuable suggestions on how to write this thesis.

This thesis would not have been possible without the support and assistance of my friends from Faculty of Science who share their valuable knowledge of thesis writing and R program running. They are so friendly and helped me to solve the problem I met.

Last but not least, I would like to dedicate this special thanks to my parents who are so concerned on my thesis progress. Thanks a lot for their support and love.

ABSTRACT

In this study, Classification and Regression Tree (CART) is used to classify and predict student who is likely to pass or fail in the final exam of Engineering Statistic course. However, two problems typical surfaced when applying CART algorithm on highly dimensional data: misclassification error and overfitting problem. Thus this research aims to reduce its misclassification error and overfitting problem for better accuracy in prediction and classification. Based on this study, different data proportion such as re-substitution method, hold-out method and 10-fold cross validation method are used for building and evaluating the decision tree. The results are compared in terms of prediction accuracy, sensitivity and specificity as well as tree structures. Based on the results obtained, 10-fold cross validation achieves the highest prediction accuracy (least misclassification error) of 85.11%. Hence, it is selected for further overfitting analysis by conducting error rate plot and cost complexity pruning methods in order to reduce the misclassification error. From the results obtained, the final pruned tree has shown to improve the prediction accuracy (87.23%). We have identified three rules generated from the final tree to identify the relationship of the attributes. Consequently, this study indicates that application of CART algorithm by 10-fold cross validation method can produce a better accuracy in classifying and predicting students' academic performance. In addition, lecturers can use such method to identify students who perform poorly in this course so that actions can be taken to avoid more failures in this course.

ABSTRAK

Dalam kajian ini, CART (*Classification And Regression Trees*) digunakan untuk mengklasifikasi dan meramal kecenderungan pelajar sama ada lulus atau gagal dalam peperiksaan akhir dalam kursus Statistik Kejuruteraan. Walau bagaimanapun, terdapat dua masalah yang biasa timbul dalam algoritma CART: ralat kesilapan dan masalah *overfitting*. Dengan itu, tujuan kajian ini dijalankan adalah untuk mengurangkan ralat kesilapan dan masalah *overfitting* demi mencapai ketepatan yang baik dalam ramalan dan klasifikasi. Berdasarkan kajian ini, pembahagian data seperti kaedah penggantian semula (*re-substitution method*), kaedah memegang keluar (*hold-out method*) dan pengesahan kaedah 10 kali ganda salib (*10-fold cross validation*) digunakan untuk membina dan menilai keputusan CART. Keputusan yang diperolehi juga dibandingkan dari segi ketepatan ramalan, sensitiviti dan spesifisiti serta struktur pokok. Berdasarkan keputusan yang diperolehi, 10 kali ganda pengesahan salib (*10-fold cross validation*) mencapai ramalan ketepatan yang tertinggi (ralat kesilapan yang terendah) iaitu 85.11%. Oleh itu, kaedah ini dipilih untuk menjalankan analisis *overfitting* dengan melukiskan kadar kesilapan dan kaedah *cost complexity* demi mengurangkan ralat kesilapan. Keputusan pokok yang dipangkas telah menunjukkan peningkatan dalam ketepatan ramalan (87.23%) dan tiga peraturan boleh dijana dari pokok itu untuk mengenal pasti hubungan atribut-atribut yang digunakan dalam kajian ini.. Oleh itu, kajian ini telah menunjukkan bahawa penggunaan algoritma CART dengan kaedah *10-fold cross validation* boleh menghasilkan ketepatan yang lebih baik dalam mengklasifikasikan dan meramalkan prestasi akademik pelajar. Di samping itu, pensyarah boleh menggunakan kaedah ini untuk mengenal pasti pelajar yang lemah dalam kursus ini supaya tindakan boleh diambil untuk mengelakkan lebih banyak kegagalan dalam kursus ini.

TABLE OF CONTENT

CHAPTER	TITLE		PAGE
	DECLARATION		ii
	DEDICATION		iii
	ACKNOWLEDGEMENT		iv
	ABSTRACT		v
	ABSTRAK		vi
	TABLE OF CONTENTS		vii
	LIST OF TABLES		x
	LIST OF FIGURES		xii
	LIST OF ABBREVIATIONS		xiv
	LIST OF APPENDICES		xv
1	**INTRODUCTION**		
	1.1	Background of Study	1
	1.2	Problem Statement	3
	1.3	Objectives	4
	1.4	Significance of the Study	4
	1.5	Scope of Study	5
	1.6	Outline of Study	5
2	**DECISION TREES IN EDUCATIONAL DATA MINIG**		
	2.1	Introduction	6
	2.2	Prediction	6
	2.3	Classification	7

	2.4	Decision Tree	10
		2.4.1 Decision Tree Mechanism	11
	2.5	A Review of Applied Decision Tree Algorithms Literature	14
	2.6	Summary	16
3	**CLASSIFCATION AND REGRESSION TREE: A REVIEW**		
	3.1	Overview of Classification and Regression Trees (CART)	17
	3.2	Basic Statistical Problem	19
	3.3	CART Decision Tree Mechanism	19
		3.3.1 Splitting Method	19
		3.3.2 Stopping Criteria	28
		3.3.3 Tree building	30
	3.4	Tree Pruning and Optimal Tree Selection	31
		3.4.1 Overfitting	31
		3.4.2 Cost-complexity	32
		3.4.3 10-fold Cross Validation	33
	3.5	Evaluation of CART	34
		3.5.1 Accuracy of Prediction	35
		3.5.2 Methods of Evaluation Accuracy	38
		3.5.2.1 Re-substitution Validation	38
		3.5.2.2 Hold-out Validation	39
		3.5.2.3 10-fold Cross Validation	39
	3.6	Summary	40
4	**DATA EXPLORATION**		
	4.1	Introduction	41
	4.2	Data	41
	4.3	Descriptive Statistics of Continuous Variables	43

	4.4	Bar Chart of Categorical Variables	45
	4.5	Data Pre-processing	47
		4.5.1 Data Discretization	47
	4.6	Research Design Framework	48
	4.6	Summary	49
5	**RESULTS AND DISCUSSION**		
	5.1	Introduction	50
	5.2	Prediction Accuracy of CART Algorithm	50
	5.3	Evaluation of Tree Structures	54
	5.4	Error Rate Plot	57
		5.4.1 Comparison of Un-pruned Tree and Pruned Tree	58
		5.4.1.1 Prediction Accuracy	58
		5.4.1.2 Tree Structure	59
	5.5	Predicting and Classifying Students' Final Exam Result	60
	5.6	Summary	61
6	**CONCLUSSIONS AND RECOMMENDATION**		
	6.1	Introduction	62
	6.2	Conclusions	62
	6.3	Recommendations	64
REFERENCES			65
Appendices A-D			69-91

LIST OF TABLES

TABLE NO.	LIST OF TABLES	PAGE
2.1	A review of applied decision tree algorithms literature	14
3.1	Example of a small training data set consists of 10 records	21
3.2	Gini index for each splitting point	23
3.3	General structure of a confusion matrix	36
3.4	Confusion matrix for CART decision tree analysis	37
4.1	A snapshot of assessment dataset obtained from Department of Mathematical Sciences, Faculty of Science, UTM	42
4.2	Variables of raw data	43
4.3	A snapshot of standardized marks of course assessment dataset	44
4.4	Descriptive statistics of continuous variables	44
4.5	UTM academic assessment is pre-classified into class “PASS” and class “FAIL”	47
4.6	A snapshot of finalised data for decision tree analysis	48

5.1	Comparison of different types of data proportions for evaluation methods of CART algorithms	52
5.2	Prediction accuracy of CART algorithms before and after pruning process	58

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	Education Data Mining Methods	8
2.2 (a)	Learning Process	9
2.2 (b)	Classification Process	9
2.3	Decision tree of student's ability in Personalised Network Learning	11
2.4 (a)	Hypothetical depth-first decision tree	13
2.4 (b)	Hypothetical best-first decision tree	13
3.1	The procedure of Classification and Regression Tree (CART)	18
3.2	CART decision tree and subset cases for a database in Table 3.1.	28
3.3	A finalised CART decision tree for database in Table 3.1	30
4.1	Bar chart of categorical variables	45
4.2	Research design framework	49
5.1	Types of data proportion	51
5.2	Tree structures of CART algorithms (before pruning)	54

5.3	Plot of re-substitution error rate (left) and the cross-validated error rate (right) for tree created on students'academic performance in Engineering Statistics Course	57
5.4	Tree structures of (a) un-pruned tree and (b) pruned tree	59

LIST OF ABBREVIATIONS

CART	-	Classification and Regression Tree
UTM	-	Universiti Teknologi Malaysia
DM	-	Data Mining

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	List of programs of Faculty of Engineering	68
B	R CODE (10-fold cross validation)	69
C	Cost complexity plot of 10-fold cross validation	76
D	A finalised data	77

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Differential Equation, Engineering Statistics, Calculus, Engineering Mathematics, and Numerical Methods are among the Mathematics courses serviced by the Department of Mathematical Sciences to students at the Engineering Faculties in Universiti Teknologi Malaysia (UTM). Over the years, the Department sometimes receive complaints from the engineering faculties on the poor performance of their students in these courses. According to UTM academic assessment system, students must at least pass the courses taken to avoid retaking the courses so that they can graduate on time.

The assessment of those courses for each student is divided into two parts: coursework assessment and final exam assessment. Coursework assessment includes Test 1(15%), Test 2 (25%) and Assignment (10%). The total marks for final exam is 50%. The marks obtained from the coursework assessment are then combined with the marks obtained from the final exam. Students who perform poorly in coursework assessment tend to result them to perform poorly in their final exam. This is because most of the chapters in Mathematics courses is tested in final exam. If students cannot catch up at the beginning chapters of these courses, they would face a big problem when they sit for their final exam. According to Kumar and Vijayalakshmi (2011), the marks obtained by students during an internal examination will play a vital role in predicting the outcome of the student in the final examination.

There are many studies done to analyse students' academic performance prior to final exam based on the classification methods in particular decision trees. For example, Yadav *et.al*., (2012) has compared three different decision tree methods: Classification and Regression Tree (CART), C4.5 and ID3 in predicting and classifying students' academic performance in order to identify which algorithms can give the highest prediction accuracy and to identify the relationship between attributes obtained in terms of classification rules. In addition, Luna (2000) has used CART to develop a classification model for predicting students' academic performance in order to determine which variables can provide information about the academic performance outcome.

Based on these related studies, it is relevant to use CART decision tree to classify students' academic performance in order to identify which coursework assessment and characteristics that most affect their final exam and predict whether the student will pass or fail in the examination. CART forms a classification tree with binary splits and has the ability to handle mixed data type (continuous and categorical variables). The performance of CART decision tree can be evaluated by identifying its misclassification error and its tree structure (i.e., the least misclassification error and the simplest tree is preferable for classification and prediction of students' academic performance).

However, two problems typically surfaced for the use of these partial dataset: misclassification error and overfitting of decision trees. Misclassification error refers to the biased estimation from a classification tree which would affect its prediction accuracy. In this study, if the misclassification error is ignored, decision tree formed is likely to misclassify students who are likely to pass or fail in final exam. Overfitting refers to the overlarge tree built consists of too many branches and replicated nodes. This phenomenon can result the generated rules from such tree to become more complicated. According to Romero *et.al.* (2011), overfitting is a critical problem in educational domain. This is because too many attributes used from a dataset can cause a formation of complex or overlarge tree with a high

misclassification error. Such occurrence yields poor prediction accuracy for future datasets.

In addition, CART decision tree methods tend to perform poorly on highly dimensional data especially the use of whole dataset as training data as well as test data. Training data refers to the available dataset used to build classification tree while test data is used to estimate the misclassification error of the training data. However, according to Du (2010) such method can cause the misclassification error called re-substitution error (the accuracy prediction is biased as the accuracy obtained is unrealistically high). This occurrence results from the same data used in building and evaluating the classification tree. One of the methods to handle this problem is to separate the data into two sets: training data for tree construction and test data for error estimation (Rokach and Maimon, 2007). Consequently, different proportions of datasets are needed for tree construction (training data) as well as tree evaluation (test data).

In this study, we use different proportion of training and test data for tree construction in order to identify which tree can produce the least misclassification error. The overfitting problem for the selected tree (least misclassification error) is thus identified in order to choose the simpler tree for classification and prediction of students' academic performance.

1.2 Problem Statement

Based on the previous sub-section, there are two CART problems typically surfaced: misclassification error and overfitting of decision tree.

Misclassification error (i.e., re-substitution error) can cause the prediction accuracy to be biased as the accuracy obtained is unrealistically high (a high but biased prediction accuracy is obtained). This occurrence results from the same data

(whole dataset) used in building and evaluating the classification tree. On the other hands, decision tree also can cause overfitting problem if the overlarge tree is built with too many branches as well as replicated nodes. Such phenomenon can result the generated rules from such tree becomes more complicated. Consequently, both of these problems would yield poor prediction accuracy for future datasets.

1.3 Objectives

1.3.1 To identify misclassification error and tree structures of CART algorithm on the whole dataset and partial dataset (training and test data) based on different proportion of datasets using Hold-out method and 10-fold cross validation method.

1.3.2 To evaluate the selected tree building based on tree size and error rate plot in order to identify overfitting problem.

1.3.3 To generate classification rules from the selected final tree and identify the relationship of the variables that most affect students' final exam result.

1.4 Significance of the Study

This study only focussed on classifying and predicting student will pass or fail in the examination by using decision tree methods. By conducting this study, lecturers can understand on how to classify and predict students' final exam result before students sit for the final exam. If the predicted outcome indicates that certain students are likely to fail in final exam, then lecturers have to put more effort to assist students to improve their performance before final exam in order to avoid them to re-take the courses in next semester.

The results of this study would give benefit to many fields (i.e., educational field, medical field and etc.) which involve in classifying and predicting mixed data (categorical and continuous attributes). Indeed, it will enhance other researchers to try other types of decision tree methods in determining classification pattern of students' academic performance.

1.5 Scope of Study

The scope of the study is to classify and predict academic performance in Engineering Statistics courses using decision trees algorithms. Engineering Statistics (course code: SSE 2193), is one of the Mathematics courses serviced by the Department of Mathematical Sciences to students at the Engineering Faculties in UTM. We have obtained a sample of 516 records of students' course assessment results and students' characteristics from three different faculties (semester 1 2009/2010).

1.6 Outline of Study

In this study, Chapter 1 discusses about the background of study, objectives, significance of study and scope of study. Chapter 2 describes general decision trees algorithms mechanism in education data mining. Next, in Chapter 3, a detailed of CART decision tree algorithm is further discussed. Chapter 4 discusses about data exploration used in this study. Chapter 5 discusses about the results and discussion. This is followed by Chapter 6 which discusses about the conclusions and recommendations for application of decision tree methods in educational data mining.

REFERENCES

Al-Radaideh, Q. A., Al-shawakfa, E. M., and Al-Najjar, M. I. (2006). Mining Student Data Using Decision Trees. *The 2006 International Arab Conference on Information Technology*. (ACIT'2006) – Conference Proceedings.

Ayesha, S., Mustafa,T., Sattar, A. R., and Khan, M. I. (2010). Data Mining Model for Higher Education System. *European Journal of Scientific Research*, 43, 24-29. EuroJournals Publishing, Inc.

Baker, R. S. J. d. (2010). *Data Mining*. Worcester Polytechnic Institute, Worcester, MA, USA: Elsevier Ltd.

Baradwaj. B. K. and Pal. S. (2011). Mining Educational Data to Analyse Students Performance. *International Journal of Advanced Computer Science (IJACSA) and Applications*, 2, 63-69.

Bienkowski, M., Feng, M., and Means, B. (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief* .U.S. Department of Education Office of Educational Technology: 9-10.

Breiman, L., J. Friedman, R. Olshen, C. Stone. (1984). *Classification and Regression Trees*. Boca Raton: Chapman & Hall.

Cai, J. (2006). *Decision Tree Pruning Using Expert Knowledge*. Doctor of Philosophy, University of Akron.

Du, H. (2010). *Data Mining Techniques and Application.* (An Introduction). United Kingdom: Course Technology.

Han, J., and Kamber, M. (2002). *Data Mining Concepts and Techniques.* USA: Morgan Kaufmann.

Han, J., Kamber, M., and Pei, J (2012). *Data Mining Concepts and Techniques.* (3rd ed.). USA: Morgan Kaufmann.

Huo, X., Kim, S. B., Tsui, K.L., & Wang, S. (2006). A Frontier-based Tree Pruning Algorithm (FBP). *Journal on Computing*, 18, 494–505. INFORMS.

Kabra, R. R. and Bichkar, R. S. (2011). Performance Prediction of Engineering Students using Decision Tree. *International Journal of Computer Applications*, 36, 8-12.

Kantardzic, M. (2003*). Data Mining: Concepts, Models, Methods, and Algorithms.* Wiley-Interscience.

Kovacic, Z. J. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. *Proceedings of Informing Science & IT Education Conference (InSITE).* Open Polytechnic, Wellington, New Zealand: InSITE, 647-665.

Kumar, S. A. and Vijayalakshmi, M. N (2011) Efficiency of Decision Trees in Predicting Students' Academic Performance. *Journal of Computer Science and Information Technology (CS & IT)*. 336. R. V. College of Engineering, Bangalore , India.

Luna, J. (2000). *Predicting Student Retention and Academic Success at New Mexico Tech*. A Master Thesis. New Mexico Institute of Mining and Technology Socorro, New Mexico.

Liu, Z., Li, H., and Zhang, Y. (2011). Application of ID3 Algorithm in Network Personalized Learning. *Energy Procedia 2011*. 9-10 December. Singapore: ESEP, 991-997.

Moertini, S. V. (2003).Towards The Use of C4.5 Algorithm for Classifying Banking Dataset. *Integral*, 8, 105-116.

Ozer, P. (2008). *Data Mining Algorithms for Classification.* BSc Thesis. Artificial Intelligence Radboud University Nijmegen.

Rokach, L. and Maimon, O. (2007). *Data Mining with Decision Trees: Theory and Applications*. London: World Scientific Publishing Co. Pte. Ltd.

Romero, C. (2011). *Handbook of Educational Data Mining.* Boca Raton: Taylor & Francis Group.

Ruggieri, S. (2009). Efficient C4.5. *Journal of Department Information, University of Pisa Corso Italy*. 2-5.

Tan, P. N., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. USA: Pearson Education, Inc.

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*. A Master Thesis. Centre of Applied Statistics and Economics Humboldt University, Berlin.

Yadav, S. K., Bharadwaj, B., Pal, S. (2012). Data Mining Applications: A Comparative Study for Predicting Student's performance. *International Journal of Innovative Technology & Creative Engineering*. 1, 13-16.

Zhao, H. (2012). The Analysis and Application of C4.5 Algorithm in Decision Tree Technology. *Advanced Materials Research*. 457-458, 754-757. Trans Tech Publication.