First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms

Azwa Abdul Aziz, Nor Hafieza IsmailandFadhilah Ahmad

Faculty Informatics & Computing
Universiti Sultan ZainalAbidin (UniSZA)
KampusTembila,
22200 Jertih, Terengganu, Malaysia
azwaaziz@unisza.edu.my

Abstract

The research on educational field that involves Data Mining techniques is rapidly increasing. Applying Data Mining techniques in an educational environment are known as Educational Data Mining that aims to discover hidden knowledge and patterns about students' behaviour. This research aims to develop Students' Academic Performance prediction models for the first semesterBachelor of Computer Science from Universiti Sultan ZainalAbidin (UniSZA)by using three selected classification methods; Naïve Bayes, Rule Based, and Decision Tree. The comparative analysis is also conducted to discover the best classification model for prediction. From the experiment, the models develop using Rule Based and Decision Tree algorithm shows the best result compared to the model develop from the Naïve Bayes algorithm. Five independent parameters(gender, race, hometown, family income,university entry mode) have been selected to conduct this study.These parameters are chosen based on prior research studies including from social sciences domains. The result discovers the race is a most influence parameter to the students' performance followed by family income, gender, university entry mode, and hometown location parameters. The prediction model can be used to classify the students so the lecturer can take an early action to improve students' performance.

Keywords: EducationalData Mining, Classification, Students' Academic Performance, Naives-Bayes

1. Introduction

Nowadays, the databases in the most organization hold so much data and information that it become complicated and difficult to analyze those data manually. To overcome the human is a limitation on handling data in the manual way, Data Mining (DM) is the suitable techniques to be used for conducting the data analysis process. Its combine machine learning, statistical and visualization techniques to discover and extract knowledge in such a way that humans can easily interpret [1, 2].

DM can be used in many areas such as medical, economic, fraud detection, engineering, and education. In recent years, there has been increasing interest in the use of DM to investigate the issues in educational field, especially at Institutions of Higher Learning (IHL) [1, 2, 3, 4]. The main objective of Educational Data Mining (EDM) is to discover the new knowledge and the hidden pattern exists in the students' data.

There is variety of DM methods that available for data experiment such as classification, clustering, and association rule [3]. This research used the classification techniques to develop a model for predicting the Student Academic Performance (SAP). This research also aims to do comparative analysis of three selected algorithms; Naïve Bayes, Decision Tree, and Rule Based to discover the best predictive model. Furthermore, the parameters that determine the SAP in Computer Science course atUniSZA are identified. The information retrieves from the model will helps the lecturer to improve the SAP and to overcome the issues of low grades obtained by students.

The six parameters are selected from the database are race, gender, family income, and university entry mode, and Grade Point Average (GPA) of the first semester of the first year BCS students. GPA is used as a dependent or the target parameter in this research and was categories into three classes; *poor*, *average*, and *good*. While the others five parameters as an independent or predictive parameter. For the experiment, the Waikato Environment for Knowledge Analysis (WEKA) open source tool is used for classification model development. WEKA is well known tool among researcher that's widely used for research purposes in the DM field [2, 4].

The rest of the paper is organized into five sections. The review of the related previous research is presented in section 2. A proposed framework for predicting SAP is described in section 3. The results of the three classification algorithms are analyzed in the result and discussion part as section 4. The paper concludes with the summary of the algorithms' performance, limitation, and further work for this research.

2. Review of Related Work

IHL faces a major challenge in order to improve and manage the educational organization to be more efficient in managing their main customer which is students in every aspect. The implementation of DM techniques in educational field is giving additional insights to the IHL in making better decisions and solutions for every issue arise [5]. The previous works show the SAP issues are the one popular subject to be explored. It's due to the demand of quality students in every public and private sector organization to fulfill the vacancies. Thus, the prediction of SAP is also important to classify the SAP so that the further action can be taken to improve students' grades.

There are various previous studies conducted to predict the SAP using DM techniques. The subsections below will present the previous research works in SAP issues and use of a DM classification techniques for prediction.

2.1 Students' Academic Performance

SAP is a main concern for every IHL including at FiC, UniSZA. By obtaining the pattern of SAP will help the students, lecturers, and administrator to take the appropriated action to improve the students' success in particular subjects or courses. The rule extracted from the prediction model also can be used to identify and classify the features of the high and low performing students. Many researches are conducted to develop the best SAP prediction model using a variety of datasets, algorithms, and tool in their case study.

The research for preventing undergraduate student retention is conducted by using 5793 records consist of six different courses which are economic sciences, law, civil engineering, languages, medicine, and pedagogy course. The 13 of 21 parameters are selected for DM during the filtering process. The experiment result will help the IHL to identify the profiles of *weak* and *excellent* students, which encompass each student for supporting their risk recommendation [6].

The educational data were analysis to improve SAP and to overcome the problem of low grades of graduate students at College of Science and Technology, Khanyonis, Gaza. The graduate students' data set consists of 3314 records and 18 parameters by classifying the grade into four groups; *excellent*, *very good*, *good*, and *average* as a target parameter. Three DM methods; association rules, classification, and clustering was applied to the data. The output from the analysis process produced a set of rules that showing the relationship between the parameters and the parameters that contributed to the SAP [7].

A research was carried out to determinants the SAP in an advanced programming course at the PETRONAS University of Technology, Perak, Malaysia. The finding of the study is able to help identifying the main indicators that may affect the students' final grade. The 24 data with four parameters tested for the analysis are coursework marks, psychosocial factors, and the information from e-learning system (total number of materials downloaded and total number of times online). The result presents the coursework marks is the most significant relationship with the final grades followed by total number of times online [8]

2.2 Classification

Classification is the one of the most commonly applied technique in EDM that predicts group exists in the data set. Classification technique is used by researchers in the educational field to better understand students' behaviors, to improve the teaching skill, and to provide the alternative solution for issues arises in IHL [9, 10, 11].

Performance analysis of engineering students is conducted using 50 data taken from different branches of the Engineering College. The data have three independent parameters which are a percentage of marks obtained in 10th class exam, percentage of marks obtained in 12th class exam, and percentage of marks obtained in Bachelor of Technology course were the percentage value are split into several classes. The final grade as a dependent parameter is divided into three categories; *excellent*, *good* and *average*. According to the experimental result from WEKA tool, the K-Nearest Neighbor (IB1) is a most suitable algorithm for the data set compared to BayesNet, Naïve Bayes, Multilayers Perceptron, Decision Table, and PART [9].

The Naïve Bayes classification algorithm is applied to the students' data at India to generate the predictive models for students' dropout management. The Naïve Bayes is applied to the 165 data consists of 17 parameters include dropout parameter that was categories into *yes* and *no* classes as a response parameter. The Naïve Bayes prediction model presents the 87% accuracy value which is the 144 from 165 records is classified correctly. The analysis result shows the male students have greater possibilities to quit the study after a year than female [10].

In another study, the three classification algorithm models are compared to discover the best model for predicting SAP to identify students at risk. Only one independent parameter used in this study, which is the midyear mark in order to predict the students of Computer Science final marks. The three algorithms were used for modeling are J48 Classifier, Decision Table, and Naïve Bayes. The J48 classifier performed better with 86% of the tuples being predicted correctly compared to 81% of accuracy for Decision Table and 84% of accuracy for Naïve Bayes [11].

From the previous research study, the three classification techniques; Naïve Bayes, Decision Tree, and Rule Based were chosen for this research. The experimentation process will be conducted using WEKA tool.

3. Proposed Framework for Predicting SAP

This section presents the proposed framework for this study in predicting SAP using DM classification techniques. The framework shows the steps in developing classification models to predict SAP of the first semester of the first year BCS students at the FIC, UniSZA [12]. Fig. 1 illustrates the Goal Identification, Data Collection, Data Pre-processing, Data Transformation, Data Mining, Result & Discussion, and Taking Action stages in this research.

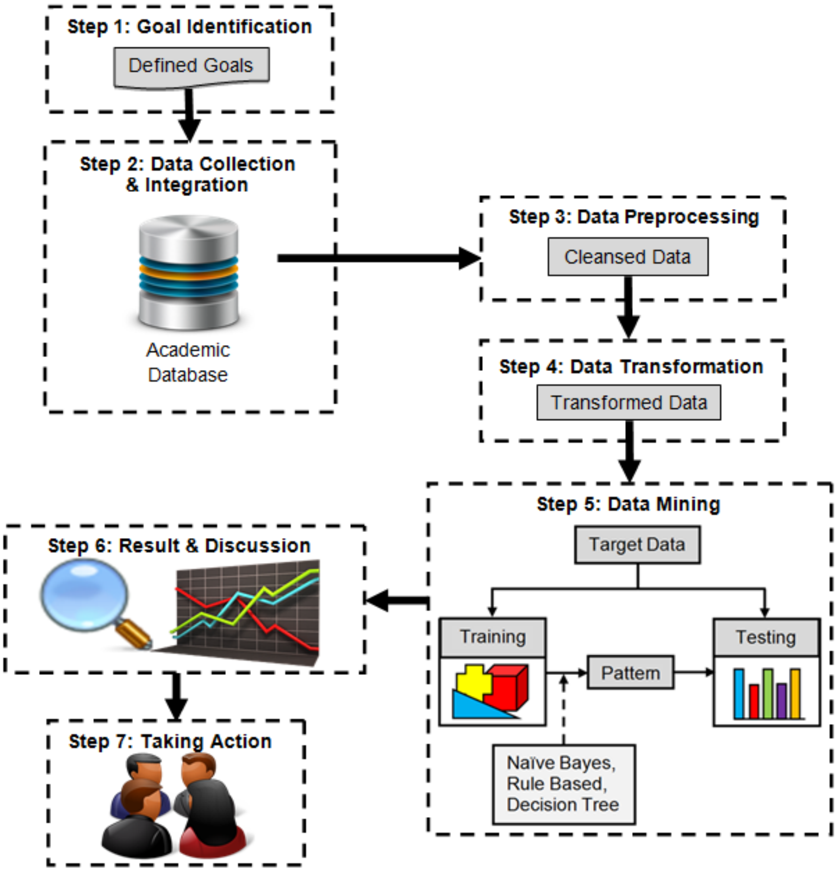
Step 1: Goal Identification
Defined Goals
Step 2: Data Collection & Integration
Academic Database
Step 3: Data Preprocessing
Cleansed Data
Step 4: Data Transformation
Transformed Data
Step 5: Data Mining
Target Data
Training
Testing
Pattern
Naïve Bayes, Rule Based, Decision Tree
Step 6: Result & Discussion
Step 7: Taking Action

Fig. 1. Proposed Framework of SAP Prediction

3.1 Goal Identification

This research goal is to develop the SAP prediction model using selected classification techniques which are Naïve Bayes, Decision Tree, and Rule Based. The WEKA open source tool written in Java language contains a collection of state-of-the-art machine learning and DM algorithms for the analysis process in this research [13]. The accuracy values of the three models are compared and the model with the highest accuracy is selected as a SAP predictive model in this case study. Additionally, the extracted knowledge from the model will be evaluated to identify the parameter that influenced the academic performance of BCS students.

3.2 Data Collection and Integration

The 497 data of BCS students, FIC, UniSZA are obtained from academic database. The data contain the information about students for a period of eight years from July 2006/2007 intakes until July 2013/2014 intakes. The data set was retrieved in two batches, the data set of the first batch contains 245 records from July 2006/2007 intakes until July 2011/2012 intakes and second batch contains 252 records from July 2012/2013 intakes until 2013/2014 intakes. The data were collected from the Academic Department of the UniSZA database that is stored in Informix Database Management System. Two separate Excel files were extracted as follows:

- Students-FIC. xlsx - This excel file contains the students' information about name, matrix number, identification number, courses' code, course's name, gender, race, religion, hometown location, type of secondary school, family income, mothers' work sector, fathers' work sector, session entry, university entry mode, Malaysian Certificate of Education grades in several subjects; Malay Language, English, and Mathematics.

- Result-FIC.xlsx - This excel file contains the students' information about matrix number, semester, GPA, and Cumulative Grade Point Average.

All the students' information is combined in a single file using the students' matrix number as a primary key during the integration process. The six parameters were selected in this research; gender, race, hometown location, family income, university entry mode, and GPA.

3.3 Data Preprocessing

The Data Pre-processing stage was performed to improve the quality of the data set by removing the missing or incomplete values in the data set. For the first batch data set, 69 records were removed from the 245 total data and only 176 records are ready for the DM process later. After the pre-processing process applied on the second batch data set, 29 records are eliminated from the data set of 252 records which left only 223 clean records. The overall, the students' data contains 98 missing values in various parameters from 497 records are ignored from the data set. The total numbers of records are reduced to 399.

3.4 Data Transformation

The Data Transformation stage will transform the numerical values into categorical values as shown in Table 1. The selected parameters are categories into two; independent and dependent parameter. Independent parameter became input of the model used in methods' equation or rules to predict the dependent parameter as an output. The independent parameters are students' gender, race, hometown location, family income and, university entry mode while a dependent parameter is GPA.

Table 1. The Transformation Selected Parameters.

Parameter Type	Parameter	Category
Independent Parameter	Gender	Male, Female
	Race	Malay, Hokkein, Chinese, Kedayan, Indian, Foochow, Bugis, Indonesian, Aboriginal People (Penisular), Brunei, Pakistani, Ghana, Semai
	Hometown Location	Town, Rural
	University Entry Mode	STPM, Diploma, Matriculation
	Family Income	No Income – 0 Low- RM1 – RM3000 Middle - RM3001 – RM7500 High - RM7501 and above
Dependent Parameter	GPA in 1st Semester of the 1st Year	Poor - 0.00 – 1.99 Average - 2.00 – 2.99 Good - 3.00 – 4.00

3.5 Data Mining

The DM stage, WEKA used to predict the SAP based on the GPA obtained in the first semester of the first year of the BCS students. The "CSV" file format was created from the Microsoft Excel to be used for analysis process. This stage consists of five phases that are training, pattern, testing, result evaluation and knowledge representation. In this stage, the cleaned or target data were divided into two parts; training and testing. The training data set used to build the model or pattern from the classification techniques and testing data set used to validate the models. After that, the result obtained will be analyzed and evaluated for further understanding about the models. The three classification algorithms are chosen for this process; Naïve Bayes, Rule Based (OneR), and Decision Tree (J48). All these algorithms will generate the predictive SAP model.

The Naïve Bayes uses the Bayes' probability theory which assumes the effect of an attribute value of a given class is independent of the values of the other attributes. It represents a descriptive and predictive approach to predict the class membership for a target tuple [14].

The Decision Tree is a powerful and a famous algorithm for classification. Decision Tree like a tree structure which start from the root of the parameter and ends with a leaf of nodes. Basically, the Decision Tree has several branches consisting of different parameter, the leaf node on each branch representing a class or a kind of class distribution. Decision tree algorithms will describe the relationships that exist among parameter in the data set. J48 implementation the C4.5 algorithm in the WEKA that creates a Decision Tree based on a set of labelled input data [15].

The Rule Based is a technique for classifying records using a collection of "IF… THEN…" rules. The OneR algorithm that was implemented in WEKA is a simple classification algorithm that generates a one-level decision tree. OneR is able to produce the simple and accurate classification rules from the data set [16].

4. Results and Discussions

The accuracy value of the model will determine how good the model can used for the prediction. Two types of data splitting used were percentages and fold cross validation. For 10:90, the training data is 10% of all data sets and 90% of all data sets used for testing data. For fold cross validation, the data sets were divided into 3, 5, or 10 subset the holdout method was repeated based on subset's number. For 3 fold cross validation, one of the 3 subset is used as the testing data and others 2 subsets were used to develop the training data. The testing data accuracy is average in 3 trails.

The first experiment is conducted using the first batch data set contains 176 clean records of six consecutive intakes from July 2006/2007 intakes until July 2011/2012 intakes. The Table 2 present the accuracy values of three classifications SAP prediction model using WEKA tool.

Table 2. The Accuracy Values of the
Classification Models for the First Experiment

		Classifiers' Accuracy		
		Naïve Bayes	**Rule Based (OneR)**	**Decision Tree (J48)**
Percentages Training : Testing	**10:90**	50.6%	50.6%	50.0%
	20:80	51.1%	47.5%	47.5%
	30:70	49.6%	44.7%	43.9%
	40:60	52.8%	49.1%	46.2%
	50:50	51.1%	51.1%	53.4%
	60:40	51.4%	54.3%	50.0%
	70:30	50.9%	54.7 %	47.2%
	80:20	54.3%	57.1%	48.6%
	90:10	61.1%	55.6%	55.6%
Fold Cross Validation	**3**	56.8%	61.9%	51.7%
	5	55.7%	57.4%	52.8%
	10	57.4%	54.5%	51.7%
Highest Accuracy		**61.1%**	**61.9%**	**55.6%**

From the result ofusing Naïve Bayes technique, its shows the best accuracy of 61.1% in 90:10 percentages test option. The Rule Based shows the best accuracy value of 61.9% in 3 fold cross validation. The highest accuracy value obtained from Decision Tree is 55.6% in test option of 90:10 percentages. The Rule Based classification model obtained highest accuracy of 61.9% compared to Naïve Bayes and Decision Tree. The detail analysis for each target class; good, average, and poor of Rule Based 3 fold cross validation are presented in confusion matrix table. This table contains the information about actual and predicted classifications using WEKA tool. Table 3 presents the confusion matrix for Rule Based of 3 fold cross validation.

Table 3. The Confusion Matrix Table for
Rule Based of 3 Fold Cross Validation

Classified as				Total of Correctly Classified Data	Prediction Success
	Poor	**Average**	**Good**		
Poor	0	5	0	0	0%
Average	0	70	22	70	76.1%
Good	0	40	39	39	49.4%

From Table 3, the model shows the better prediction for average category with 76.1% of accuracy, 49.4% of accuracy for good category is correctly classified and 0% of accuracy for poor category. This prediction model gives a better classification for average student category and failed to predict the poor student category.

The experiment continues by adding more records from second batch data of two consecutive intakes from July 2012/2013 and July 2013/2014. 223 clean records were combined in the same file of 176 records from the first batch data and the end total of records for ready the next experiment is 399.
The Table 4 shows the accuracy value obtained from WEKA using 399 records. The data set is applied using the same three classification algorithms as the previous experiment.

Table 4. The Accuracy Values of the
Classification Models for the Second Experiment

		Classifiers' Accuracy		
		Naïve Bayes	**Rule Based (OneR)**	**Decision Tree (J48)**
Percentages Training : Testing	**10:90**	49.3%	49.0%	44.3%
	20:80	47.6%	48.6%	44.5%
	30:70	49.1%	45.5%	43.0%
	40:60	49.4%	49.0%	49.8%
	50:50	50.8%	48.7%	42.2%
	60:40	52.5%	50.6%	51.9%
	70:30	63.3%	64.2%	64.2%
	80:20	58.8%	68.8%	68.8%
	90:10	60.0%	62.5%	62.5%
Fold Cross Validation	**3**	50.6%	53.9%	51.9%
	5	52.6%	54.6%	53.1%
	10	50.9%	54.4%	51.4%
Highest Accuracy		**63.3%**	**68.8%**	**68.8%**

From Table 4, the Naïve Bayes shows the best accuracy of 63.3% in 70:30 percentages test option. The Rule Based and Decision Tree shows the best accuracy value of 68.8% in 80:10 percentages. It shows the Rule Based and Decision Tree is the best classification model for the second experiment. Table 5 presents the confusion matrix for the model extracted from Rule Based and Decision Tree algorithms in 80:20 percentages test option.

Table 5. The Confusion Matrix Table for
Rule Based and Decision Tree of 80:20 percentages

Classified as				Total of Correctly Classified Data	Prediction Success
	Poor	**Average**	**Good**		
Poor	0	2	0	0	0%
Average	0	49	0	49	100%
Good	0	23	6	6	20.7%

The Table 5 shows the models predict success of 100% in average category, followed by good category with 20.7% of predictions success. But, the models failed to predict poor category.

The feature selection is conducted by applying Information Gain (InfoGain) calculation of the parameters using WEKA. InfoGain used to discover which parameter that influences the SAP. InfoGain parameter evaluation involves Entropy calculation, which the InfoGain(Class,Parameter) = Entropy(Class) – Entropy(Class | Parameter). Table 6 presents the InfoGain for each parameter used in this research.

Table 6. The Information Gain for Five Dependent Parameters

Parameter	Information Gain
Race	0.07090
Family Income	0.02399
Gender	0.01409
University Entry Mode	0.01019
Hometown Location	0.00204

The race parameter shows the highest InfoGain value of 0.07090 bits, it shows that race is a most influencing parameter for the SAP in this case study. The InfoGain for the family income parameter is 0.02399 bits, gender parameter is 0.01409 bits, and university entry mode parameter is 0.01019 bits. The hometown location parameter gives the less influence of students' success.

5. Conclusion

The amount of data stored in an educational database at HLI is increasing rapidly by the times. In order to get the knowledge about student from such large data and to discover the parameter that contributed to the students’ success, the classification techniques are applied to the students’ data. The result reveals the models of Rule Based and Decision Tree algorithm gives the highest prediction accuracy value of 68.8%. The show the excellent performance in predicting average students with the accuracy value of 100%, but the model failed to predict the poor students. The models’ performance is increased with the increase of the number of records for analysis. It's concludes that the model is becoming more accurate for prediction if big data is involved in DM process. The limitation of this study is the small size of data due to incomplete and missing value. For the next experiment, more data will be added so the accuracy of prediction model can be improved for FIC community benefits in monitoring the SAP.

References

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). *Use Data Mining To Improve Student Retention In Higher Education – A Case Study*. ICEIS - 12th International Conerence on Enterprise Information Systems, 190–197.

Pal, S. (2012). *Mining Educational Data Using Classification to Decrease Dropout Rate of Students*.International Journal of Multidisciplinary Sciences and Engineering, 3(5), 35–39.

Baradwaj, B. K. (2011). *Mining Educational Data to Analyze Students' Performance.(IJACSA)* International Journal of Advanced Computer Science and Applications, 2(6), 63–69.

Kabakchieva, D. (2013). *Predicting Student Performance by Using Data Mining Methods for Classification*.Cybernetics and Information Technologies, 13(1), 61–72. doi:10.2478/cait-2013-0006.

Prakash, S., Ramaswami, K. S., & Post, C. A. (2010*).Fuzzy K- Means Cluster Validation for Institutional Quality Assessment*.Communication and Computational Intelligence (INCOCCI), 2010 International Conference, 628–635.

Silva, H. R. B. da, &Adeodato, P. J. L. (2012).*A data mining approach for preventing undergraduate students retention*. The 2012 International Joint Conference on Neural Networks (IJCNN), 1–8. doi:10.1109/IJCNN.2012.6252437

Tair, M. M. A., & El-Halees, A. M. (2012).*Mining Educational Data to Improve Students' Performance : A Case Study*. International Journal of Information and Communication Technology Research, 2(2), 140–146.

Chen, Y. Y., Taib, S. M., Sarah, C., &Nordin, C. (2012).*Determinants of Student Performance in Advanced Programming Course*.The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012) Determinants, 304–307.

Singh, S., & Kumar, V. (2013*).Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques*.International Journal of Computer Science Engineering and Technology (IJCSET), 3(2), 31–37.

Pal, S. (2012*). Mining Educational Data Using Classification to Decrease Dropout Rate of Students.International Journal of Multidisciplinary Sciences and Engineering*, 3(5), 35–39.

Mashiloane, L., &Mchunu, M. (2013). Mining for Marks : *A Comparison of Classification Algorithms when Predicting Academic Performance to Identify " Students at Risk ." In Lecture Notes in Computer Science (pp. 541–552).*

Aziz, A. A., Ismail, N. H., & Ahmad, F. (2013).*Mining Students' Academic Performance.Journal of Theoretical and Applied Information Technology*, 53(3), 485–495.

Wahbeh, A. H., Al-radaideh, Q. A., Al-kabi, M. N., & Al-shawakfa, E. M. (2011).*A Comparison Study between Data Mining Tools over some Classification Methods*.(IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, 18–26.

Kumar, U., & Pal, P. S. (2011). Data Mining : *A prediction of performer or underperformer using classification*. International Journal of Computer Science and Information Technologies (IJCSIT), 2(2), 686–690.

Gholap, J. (2012). *Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility*. Asian Journal of Computer Science and Information Technology, 2(8).

Buddhinath, G., & Derry, D. (2006).*A Simple Enhancement to One Rule Classification. Retrieved* from http://www.buddhinath.net/OtherLinks/Documents/Improved OneR Algorithm.pdf