Journal of Engineering and Applied Sciences 13 (Special Issue 6): 5302-5307, 2018
ISSN: 1816-949X

A Review of Student's Performance Prediction Using Educational Data Mining Techniques

Annisa Uswatun Khasanah
Department of Industrial Engineering, Universitas Islam Indonesia, Yogyakarta, Indonesia

Abstract: The educational institutions can use educational data mining to analysis student's performance which can help the institution in identifying the student fail or dropouts. Classification is popular data mining technique that have been widely implemented to predict student performance. There are many attributes and methods that can be used to predict student's performance. This preliminary study presents study review related with this current topic to reveal the most widely used classification technique and the attributes that commonly used to predict student performance. Qualitative content analysis is used as the research method. It can be concluded that the most widely used classification methods based on the literature reviews are decision tree and Bayesian network and method that outperform the other based on the comparison analysis from several studies is decision tree with CART algorithm while the most widely used attributes can be categorized into four categories, student personal information, family information, pre-university characteristics and university features.

Key words: Paper review, student performance, prediction, educational data mining, comparison, classification

INTRODUCTION

There was a rapid increase in the number of college, since, 2005 in Indonesia. Chairman of the Association of Indonesian Private University, Edy Suwandi Hamid, said that in this last ten years one new college was established every 2 days. In 2005 there are 2408 colleges in Indonesia but now this number is doubled into 4264 (Anonymous, 2015). While in Yogyakarta which is popular as a student city in 2015 there were 130 colleges including academy, polytechnic, institute, university and other (Anonymous, 2016).

This current condition leads to high competition among the educational institutions. They all promise to provide quality education to its students. Student are the main asset for universities and none of them would like to miss opportunity to let new students come every year. The student performance plays important rule to measure the quality of a university. The university must concern about the performance of the students to produce the best quality of graduates. Every university graduate aims to get a prospect job after finishing their study and it cannot be ignored that currently academic performance is one of the main factors considered by the employer in recruiting workers, especially, the fresh graduates (Baradwaj and Pal, 2011a, b; Yadav *et al*., 2012). Predicting the student performance is one way that can be conducted by university to monitor their student and to prevent student failed.

Data mining is one of popular method that have been widely used by many scholars to predict student performance. Data mining is the process of automatically discovering useful information in large data repositories (Tan *et al*., 2006). Nowadays, there are increasing research interest in using data mining in education and data mining technique that applied in this filed can be usually called as education data mining. Yadav and Pal (2012). Data mining provides many tasks including classification that can be implemented to predict student performance.

Classification is one of popular data mining technique that have been widely used in many different scopes of study. Classification can be defined as the task of assigning objects to one of several predefined categories (Tan *et al*., 2006). It employs a set of pre-classified example to build a model that can classify the population of large records (Ahmed and Elaraby, 2014). In educational scope, classification have been widely implemented to predict student's performance. There have been many scholars that implement classification technique to conduct student's performance prediction. Some scholar define performance using the final score of an exam, fail or pass in a specific course, drop out or not and Grade Point Average (GPA) as the parameter. They use different method such as decision trees, neural networks, Naive Bayes, k-nearest neighbor and many others. They also used different attributes to predict the class data. To find out suitable attributes and good classification methods are essential to get powerful

prediction model. This study is a review about student's performance prediction by collecting related research to reveal techniques and also, attributes that have been commonly used in this issue.

By reviewing related studies, the objectives of this study are to know the most widely used classification technique and the attributes that commonly used to predict student's performance. This current study is preliminary study and the output will be considered for the further proposed study. The rest of this study is organized as follows.

Literature review: Educational data mining is a new field in data mining to explore educational data to determining the usefulness of learning systems, analysis learner academic performance and developing an early warning system. The educational institutions can use educational data mining to analysis student's performance which can help the institution in identifying the student fail or dropouts and allow the teacher to provide appropriate counselling. There have been many studies that conducted by many scholars related with this issues (Al-Radaideh *et al*., 2006; Ahmed and Elaraby, 2014; Baradwaj and Pal, 2011a, b; Kabakchieva, 2013; Kumar and Vijayalakshmi, 2011; Mueen *et al*., 2016; Ramaswami and Rathinasabapathy, 2012; Yadav *et al*., 2012; Yadav and Pal, 2012).

Mueen *et al*. (2016) conducted research using educational data mining to predict student's performance using three different data mining classification algorithms (decision tree, neural network using multilayer perception with back-propagation type supervised-learning algorithm and Navie Bayes). The prediction performance of three classifiers are measured and compared using data set from undergraduate students who had taken the programming fundamental and advanced operating system courses from August 2014-May 2015. The attributes that used in this study were categorized into three types, general, forum and academic. The data analysis was performed in software WEKA. Pre-processing data, including feature selection, imbalanced data and data transformation was conducted before applying the classification techniques. Feature selection was conducted to identify attributes which have greater impact. And it can be revealed that some features are not related to student performance. From the 38 available attributes, there are 7 best attributes that selected by more algorithms, including, GPA, two lab test assignment submit, rate of forum join, attendance, average of lab test and grade of final exam. Imbalanced data analysis was conducted in the dataset after feature selection. Data is imbalanced when number of instances in one class is much smaller than the number of instances in other class. Data transformation was used to integrate the data obtained from different sources into one single dataset. The dataset consisted of 60 students where 41 passed 68.33% and 19 failed (31.66) of the course final exam. To evaluate and compare classifier performance (Mueen *et al*., 2016) used accuracy, precision, recall and specificity. Accuracy rate which represent the effectiveness of the classifier shows Navie Bayes performs better than the other two. Navie Bayes is also, the winner in precision which shows the predicative power. According to recall which represents the sensitivity, multilayer perception performs better. In specificity again Navïe Bayes outperforms others. But overall, the results show that Navie Bayes classifier outperforms other two classifiers by obtaining the overall prediction accuracy of 86%.

Ahmed and Elaraby (2014) also tried to implement educational data mining to conduct student's performance analysis. The decision tree (using ID3 algorithm) method was used in this study. The dataset was obtained from a student's database used in one of the educational institutions on the sampling method of Information system department from session 2005-2010 and initially size of the data is 1547 records. The attributes that included in this study were student department, degree of high school, marks of midterm, grade of lab test, seminar performance assignment, measure of student participate, homework and the target class is final grade mark. Midterm mark has the highest gain, then it was used as the root node. Ahmed and Elaraby (2014) conlcuded that by conducting student performance analysis can help the student's to improve the student's performance. In addition it can identify students which needed special attention to decrease students failed and take a right action at the appropriate time.

Kabakchieva (2013) presented the initial results from a data mining research project implemented at a Bulgarian university. The main goal of the research aimed at revealing the high potential of data mining applications for university management. The specific objective of the proposed research work by Kabakchieva (2013) was to find out if there are any patterns in the available data that could be useful for predicting student's performance at the university based on their personal and pre-university characteristics. Decision tree classifier, Bayes classifiers and a nearest neighbour classifier were the classification methods that used to analyze the data. The record included 10330 students and 14 attributes. The class attribute used numeric parameter of university average score including "excellent", "very good", "good", "average" and "bad". Popular WEKA classifiers were used in the experimental study, including a common

decision tree algorithm C4.5 (J48), two bayesian classifiers (NaiveBayes and BayesNet), a nearest neighbour algorithm (IBk) and two rule learners (OneR and JRip). The output show that the decision tree classifier (J48) performs best, since, it has highest overall accuracy, followed by the rule learner (JRip) and the kNN classifier. The Bayes classifiers are less accurate than the others. However, all tested classifiers are performing with an overall accuracy below 70% which means that the error rate is high and the predictions are not very reliable.

Ramaswami and Rathinasabapathy (2012) used bayesian network approach in the field of education to predict student's performance. The data contained 35 attributes with 5650 records and for the classattribute HSc level grade was used. The feature selection as well as student performance prediction models are extensively studied by varying the number of cases of class attribute, two-case (pass, fail), three-case (very-good, good, poor), five-case (excellent, very-good, good, fair, poor) and seven-case values (O, A-F). The data analysis for this study was conducted using sofware WEKA. This study used a simple select estimator algorithm, like hill Climbing, K2, LAGD hill Climbing, Repeated hill Climbing, Tabu search and network augmented with Tree to find the conditional probability tables of the Bayes network. The result showed that Bayesian network models with Network Augmented with Tree search algorithm achieves better performance over others types of search algorithms for all types of class values. The result also revealed that the class attribute HScGrade has strong dependency on marks obtained at secondary level, type of transportation to school, medium of instruction, sibling structure (number of brothers and sisters) and economic status of the family.

Other scholars, Yadav and Pal (2012), implemented educational data mining techniques to build a prediction model for engineering student's performance. The C4.5, ID3 and CART decision tree algorithms were applied to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to pass, fail or promoted to next year. In this study the dataset was obtained from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method for Institute of Engineering and Technology for session 2010. Initially size of the data is 90. The data was performed in software WEKA. The result showed eventhough C4.5, ID3 and CART algorithms showed an acceptable level of accuracy, C4.5 technique has highest accuracy of 67.7778% compared to other methods.

Like the previous scholars (Yadav *et al*., 2012) implemented decision tree algorithms (C4.5, ID3 and CART) to predict student's performance. The data set used in this study was obtained from VBS Purvanchal University, Jaunpur (Uttar Pradesh), India on the sampling method of computer applications department of course Master of Computer Applications from session 2008-2011. Initially size of the data is 48 records. And the attribute that used in this study were previous semester marks, class test grade, seminar performance assignment, attendance, lab research and for the class attribute was end semester marks (first, second, third, fail). The result indicated that a CART technique has highest accuracy of 56.25% compared to other methods. ID3 algorithm also showed an acceptable level of accuracy.

Bhardwaj and Pal (2011a, b) only used ID3 decision tree algorithm to predict student's performance. The data set used in this study aldo was obtained from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method of computer applications department of course master of computer applications but from different session. In this study the data was taken from session 2007-2010 and the initially size of the data was 50. Most attribute that used in this study were the same with the previous study but in this study Bhardwaj and Pal (2011a, b) also used general proficiency.

Different with the previous one in this study Baradwaj and Pal (2011a, b) tried to imp lement Bayesian Network to predict student's performance. The dataset was collected different degree colleges and institutions affiliated with Dr. R. M. L. Awadh University, Faizabad of Computer Applications Department of course Bachelor of Computer Applications year 2009-2010 and the initial data was 300 (226 males, 74 females). In this study Baradwaj and Pal (2011a, b) used more attributes than the previuos one, including student gender, student category, teaching language, student food habit, student other habit, living location, size of family, status of family, family annual income, grade in senior secondary education, college type, qualification (father and mother), occupation (father and mother) and for the class attribute, the final mark was used (first, second, third, fail). It was found that the student's performance was highly dependent on their grade obtaines in senior secondary examination, living location and the teaching language.

Kumar and Vijayalakshmi (2011) implement educational data mining to predict student's performace using decision tree with C4.5 algorithm, ID3 algorithm like (Yadav and Pal, 2012; Yadav *et al*., 2012). The algorithms were implemeted in WEKA. This study try to predict wheater the student will pass or fail in the examination based on the marks obtained by the students during the internal examination (MCA11-MCA15). The dataset consisted of 100 data where 39 indicated fail and 60 indicated pass. The result indicated that C4.5 algorithm was more accurate than ID3 algorithm.

Al-Radaideh *et al.* (2006) tried to predict the performance of Technology and Computer Science Faculty, Yarmouk University, Jordan students who took the Programming I course (C++). This study also, tried to compare ID3 algorithm, C4.5 algorithm and Naive Bayes. There were 12 attributes, included gender, age, department, high school major, high school grade, study type, funding, place of residency, lecturer degree, lecturer gender, lecturer department, number of repetition and the grade of the C++ course (A-D). WEKA was used to analize the data. The high school grade had the highest gain ration then this attribute was considered as the root node of the decision tree. The result showed that the classification accuracy for the three different classification algorithms was not so high. It indicated that the collected samples and attributes were not sufficient to generate a classification model of high quality.

MATERIALS AND METHODS

In this study, some journal articles related with student's performance prediction were analyzed using qualitative content analysis methodology. Content analysis is a qualitative method to summarize any form of content from various aspect of resources. This analysis aims to compare previous reserach and find out the most widely used attributes and classification method to predict student performance. The journal study were collected from Google Scholar using keyword student's performance, prediction and data mining. About 30 journal articles were collected. The 30 journal study were skim read to find which study compared some classification method in predicting student's performance, then 10 journal study were selected to be analyzed. Scan read of the 10 journal articles were conducted to find some aspect including attributes, class attribute, methods, application software and the results.

RESULTS AND DISCUSSION

In this study, the research objectives to know the most popular classification method and also, the most widely used attributes to predict student's performance will be discussed. The output from this preliminary study will be considered in the further research.

Popular classification method to predict student's performance: Several studies related with the implementation of educational data mining to predict student's performance have been discussed in the study 2. From those discussions it can be concluded that the most widely used classification is decision tree (Ahmed and Elaraby, 2014; Kabakchieva, 2013; Mueen *et al.*, 2016; Yadav *et al.*, 2012; Yadav and Pal, 2012; Bhardwaj and Pal, 2011a, b; Kumar and Vijayalakshmi, 2011), bayesian network (Al-Radaideh *et al.*, 2006; Baradwaj and Pal, 2011; Kabakchieva, 2013; Mueen *et al.*, 2016; Ramaswami and Rathinasabapathy, 2012), Neural Network (Mueen *et al.*, 2016) and other (Kabakchieva, 2013).

For the decision three algorithm, some scholars tried to compare the performance of ID3, C4.5 and CART (Al-Radaideh *et al.*, 2016; Kumar and Vijayalakshmi, 2011; Yadav *et al.*, 2012; Yadav and Pal, 2012). Each of the scholar have different conclusion according with which algorithm outperform the other. Yadav and Pal (2012) and Yadav *et al.* (2012) concluded that CART algorithm indicated have better result than ID3 and C4.5 based on the accuracy rate. Kumar and Vijayalakshmi (2011) had different conclusion, the result indicated that C4.5 algorithm was outperform the other algorithms. While (Al-Radaideh *et al.*, 2006) concluded that none of the algorithms showed a good result because of the insufficient data.

Form the previous discussion there also some scholars that tried to compare the performance of decision tree, bayesian network and other classification method like multi layer perceptron and nearest neighbor (Al-Radaideh *et al.*, 2006; Kabakchieva, 2013; Mueen *et al.*, 2016). Mueen *et al.* (2006) concluded that Bayesian network in this case was Naive Bayes was better than decision tree and neural network (multi layer perceptron). While Kabakchieva (2013) and Al-Radaideh *et al.* (2006) had different point of view, they concluded that decision tree better than Bayesian network and also nearest neighbour method. Then from the whole analysis it can be conclude that the most powerful classification based on the literature review is decision tree with CART algorithm. It also can be concluded that decision trees are so, popular because they produce classification rules that are easy to interpret than other classification methods (Bhardwaj and Pal, 2011a, b; Yadav *et al.*, 2012; Yadav and Pal, 2012).

Popular attributes to predict student's performance: From all related studies that have been discussed there were no specific conclusion what kind of attributes that good or bad to be the parameter in student' performance prediction. The attributes that used by scholars that have been mentioned in study 2 can be categorized into 4 categories: student personal information, family information, pre-university characteristic and university feature. Table 1 and 2 show the attribute categories.

The attributes that mentioned in Table 1 are the popular attributes used by scholars as the parameters to predict student's performance. In the initial phase, some scholars used many attributes and conduct the feature

Table 1: Attributes that commonly used as the parameter to predict student's performance

Category	Attributes
Personal information	Gender
	Age
Family information	Father education
	Father occupation
	Mother education
	Mother occupation
	Family size
	Family income
Pre university characteristic	The final secondary education score
	Total admission score
	Profile secondary education
University features	Attendance
	Final exam grade
	Lab test
	GPA

Table 2: Attributes used for the proposed study

Category	Attributes
Personal information	Gender (GE)
	Origin (OR)
Family information	Father Education (FE)
	Father Occupation (FO)
	Mother Education (ME)
	Mother Occupation (MO)
Pre-university characteristic	Senior high school Type (ST)
	Senior high school Department (SD)
	Senior high school Final grade (SF)
University features	First semester Attendance (AT)
	Final GPA (FGPA)
	Drop out or Not (class attribute)

selection to reduce the number attributes. Then, only important attributes that were concluded for the further analysis. Bharadwaj and Pal (2011a, b) used 17 attributes in the initial phased. Then after doing attribute selection using MATLAB only 8 high influencing attributes included in the analysis.

Ramaswami and Rathinasabapathy (2012) used three type of feature selection to evaluate the high influencing attributes included, Consistency subset Evaluation (CFS), chi-square based attribute evaluation (CHI) and Information Gain Attribute Evaluation (ING). Ramaswami and Rathinasabapathy (2012) concluded that information Gain attribute evaluation (ING) indicated better performance that the other type. In the initial step they used 35 attributes in used the HScGrade (Higher Secondary Grade) as the class label and used varying number of class, two-case (pass, fail), three-case (very-good, good, poor), five-case (excellent, very-good, good, fair, poor) and seven-case values (O, A-F). From the feature selection in can be concluded that the ING method indicated 9 top ranked attributes for two class, 13 top ranked attributes for three class, 19 attributes for five class and 23 attribute for seven class. In other study (Mythili and Shanavas, 2014) concluded that the student's academic performance was influenced by various factors like parent's education, locality, economic status, attendance and gender.

For the popular class attributes that used in student's performance prediction, scholars used different kinds of attributes based on the purpose of the study. Mueen *et al.* (2016), Yadav and Pal (2012) and Kumar and Vijayalaksmi (2011) used pass or fail in a specific course as the class attributes. The other scholars used more than one class to determine the final grade of the students as the class attribute (Al-Radaideh *et al*., 2006; Bhardwaj and Pal, 2011a, b; Kabakchieva, 2013; Ramaswami and Rathinasabapathy, 2012; Yadav *et al*., 2012).

CONCLUSION

From this study, it can be concluded that the most widely used classification methods based on the literature reviews are decision tree and Bayesian network and method that outperform the other based on the comparison analysis from several studies is decision tree with CART algorithm. There is no specific conclusion related with the best attributes as the parameter that can be used for predicting student's performance. From literature study, the most widely used attributes can be categorized into four categories, student personal information, family information, pre-university characteristics and university features.

SUGGESTIONS

The results from this current study is very important for the further proposed study. The information related with the popular classification method and the common attributes used to predict student's performance will be considered. In the further proposed research, there will be 13 attributes as shown in this study. The data are collected from student data base that can be accessed from the information system of Universitas Islam Indonesia (UNISYS). The data consist of 104 data of 2007 student. There are 13 data 12.5% classified as DO and 91 data 87.5% classified as not. The data collection is still in progress. All of the data will be used in the analysis. The data will be divided become training data to build the classification model and test data to validate the model. Decision tree using CART algorithm and Bayesian network will be used as these two methods are the most popular method based on the literature review. Accuracy rate will be used to validate the classification model.

REFERENCES

Ahmed, A.B.E.D. and I.S. Elaraby, 2014. Data mining: A prediction for student's performance using classification method. World J. Comput. Appl. Technol., 2: 43-47.

Al-Radaideh, Q.A., E.M. Al-Shawakfa and M.I. Al-Najjar, 2006. Mining student data using decision trees. Proceedings of the 2006 International Arab Conference on Information Technology (ACIT'06), December 19-21, 2006, Yarmouk University, Irbid, Jordan, pp: 1-5.

Anonymous, 2015. [Every two days, one college appears in Indonesia]. Tempo Inti Media, Jakarta, Indonesia. (In Indonesian) https://nasional.tempo.co/read/1109898/tgb-zainul-majdi-mundur-dari-demokrat-karena-alasan-pribadi

Anonymous, 2016. [Graph number of universities]. Ministry of Research, Technology and Higher Education, Jakarta, Indonesia. https://translate.google.com/translate?hl=en&sl=id&u=https://forlap.ristekdikti.go.id/perguruantinggi/homegraphpt&prev=search

Bharadwaj, B.K. and S. Pal, 2011b. Data mining: A prediction for performance improvement using classification. Int. J. Comput. Sci. Info., Security (IJCSIS), 9: 136-140.

Bharadwaj, B.K. and S. Pal, 2011a. Mining educational data to analyze student's performance. Int. J. Adv. Comput. Sci. Applic., 2: 63-69.

Kabakchieva, D., 2013. Predicting student performance by using data mining methods for classification. Cybern. Inf. Technol., 13: 61-72.

Kumar, S.A. and M.N. Vijayalakshmi, 2011. Efficiency of decision trees in predicting student's academic performance. Comput. Sci. Inf. Technol., 1: 335-343.

Mueen, A., B. Zafar and U. Manzoor, 2016. Modeling and predicting students academic performance using data mining techniques. Intl. J. Mod. Educ. Comput. Sci., 8: 36-42.

Mythili, M.S. and A.M. Shanavas, 2014. An analysis of students' performance using classification algorithms. IOSR. J. Comput. Eng., 16: 63-69.

Ramaswami, M. and R. Rathinasabapathy, 2012. Student performance prediction. Intl. J. Comput. Intel. Inf., 1: 231-235.

Tan, P.N., M. Steibach and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley, Boston, MA., USA., ISBN-13: 9780321420527, Pages: 769.

Yadav, S.K. and S. Pal, 2012. Data mining: A prediction for performance improvement of engineering students using classification. World Comput. Sci. Inf. Technol. J., 2: 51-56.

Yadav, S.K., B.K. Bharadwaj and S. Pal, 2012. Data mining applications: A comparative study for predicting student's performance. Int. J. Innovative Technol. Creative Eng., (IJITCE), 1: 13-19.