http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (4): 832-837, 2010
ISSN 1812-5638

Ranking of Influencing Factors in Predicting Students' Academic Performance

L.S. Affendey, I.H.M. Paris, N. Mustapha, Md. Nasir Sulaiman and Z. Muda
Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

Abstract: The aim of this study was to rank influencing factors that contribute to the prediction of students' academic performance. It is useful in identifying weak students who are likely to perform poorly in their studies. In this study, we used WEKA open source data mining tool to analyze attributes for predicting a higher learning institution's bachelor of computer science students' academic performance. The data set comprised of 2427 number of student records and 396 attributes of students registered between year 2000 and 2006. Preprocessing includes attribute importance analysis. We applied the data set to different classifiers (Bayes, trees or function) and obtained the accuracy of predicting the students' performance into either first-second-upper class or second-lower-third class. A cross-validation with 10 folds was used to evaluate the prediction accuracy. Our results showed the ranking of courses that has significant impact on predicting the students' overall academic results. In addition, we perform experiments comparing the performance of different classifiers and the result showed that Naïve Bayes, AODE and RBFNetwork classifiers scored the highest percentage of prediction accuracy of 95.29%.

Key words: Predicting academic performance, data mining, attributes ranking

INTRODUCTION

Predicting student performance in an academic program is a difficult but useful undertaking. Most higher learning institutions have systems to store student's information and these databases contain useful knowledge that can be extracted. For this reason, we embarked on a project to extract useful information from the student information system of a higher learning institution which is believed to contain a wealth of detailed information about performance indicators. However, having looked at the data, there was insufficient background information about the students' academic entry qualification to co-relate with their actual academic performance. In addition, due to confidentiality and sensitive issues, attributes such as gender and race are not allowed to be used. With the small and sparse data set that we have, we decided to extract the attribute importance ranking of courses to determine which courses have significant contribution to the prediction of the overall academic performance. The predictive task can be achieved by using data mining tools. We group the courses into three different components, namely, core courses (computer science, mathematics), university courses and elective courses. In addition, we distinguished courses as first-year, second-year or third-year courses. This information is very useful for the Faculty's management to monitor the deliverables of the top ranking courses.

Curriculum committees can use prediction results to guide changes to the curriculum and evaluate the effects of those changes. An academic advisor can refer to the prediction results when giving advice to students who perform weakly in their studies so that preventive measures can be taken much earlier. In addition, an instructor can further improve his/her teaching and learning approach, as well as plan interventions and support services for weak students.

Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of future observations, such as predicting the students' academic performance. Data mining techniques can discover useful information that can be used in formative evaluation to assist educators establish a pedagogical basis for decisions when designing or modifying an environment or teaching approach.

Many studies on educational data mining made used of data from web-based educational systems which record students' accesses in web logs (Minaei-Bidgoli *et al*., 2003; Merceron and Yacef, 2005;

Corresponding Author: Lilly Suriani Affendey, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
Tel: +60-3-8946-6549 Fax: +60-3-8946-6577

Romero *et al.*, 2008). Most of these researches are is to provide adaptation to a learner using the data stored in the student model. The patterns of use in the data gathered are used to make predictions as to the most-beneficial course of studies for each learner based on their present usage.

This study on the other hand, use historical data of students' academic results and the main objective was to rank courses using attribute importance analysis and to predict whether, a first year student will graduate higher or lower than a second class upper. A second goal was to compare the prediction accuracy of several classification methods.

Many studies have attempted to predict students' academic performance in order to enhance teaching and learning. Various techniques include statistical analysis, machine learning and data mining.

Chamillard (2006) used statistical analysis techniques to predict student performance in a particular course. Observations from the analysis provide useful insight into the relationships between courses.

Superby *et al.* (2006) and Vandamme *et al.* (2007) studied correlations of various parameters such as attendance, estimated chance of success, previous academic experience and study skills. They found out that changing process factors during a student's stay at the university plays a large part in academic performance. In addition they experimented on predicting students' performance using decision tree, neural networks and linear discriminant analysis. The rates of prediction obtained were not particularly good due to the difficulty to classify students into 3 groups, namely, high risk, medium risk and low risk, before the first university examinations.

Golding and Donaldson (2006) stated that the use of performance in first year computer science course is a possible factor, which may determine academic performance. They also showed that gender and age have no significant correlation as predictive factors.

McKenzie and Schweitzer (2001) investigated academic, psychosocial, cognitive and demographic predictors of academic performance to improve interventions and support services for student at risk of academic problems. They recommended implementing stringent record keeping procedures at the university level to enable researchers to fully examine the relationship between age, previous academic performance and university achievement.

Merceron and Yacef (2005) made used of data mining algorithms to discover patterns that are pedagogically interesting. Their findings were used to help teachers manage their class, understand their students' behavior and support learner reflection through proactive feedback. Their data set is collected from a web-based tutoring tool focusing on logic proofs exercises.

Kotsiantis *et al.* (2003) predicted drop-outs in the middle of a course by comparing six classification methods (Naive Bayes, decision tree, feed-forward neural network, support vector machine, 3-nearest neighbor and logistic regression). The data set which consisted of 350 records contained demographic data, results of the first writing assignments and participation to group meetings. Their best classifiers, Naive Bayes and neural network, were able to predict about 80% of drop-outs.

Minaei-Bidgoli *et al.* (2003) predicted the course final results from a learning system log data by comparing six classifiers (quadratic Bayesian classifier, 1-nearest neighbors, k-nearest neighbors, Parzen window, feed-forward neural network and decision tree). The data which consisted of 250 records contained attributes concerning each task solved and other actions like participating in the communication mechanism and reading support material. Their best classifier, k-nearest neighbors, achieved over 80% accuracy, when the final results had only two classes (pass/fail).

Minaei-Bidgoli *et al.* (2004) applied data mining classifiers as a means of analyzing and comparing use and performance of students who have taken a technical course via the web. Their results show that combination of multiple classifiers leads to a significant accuracy improvement in the given data set.

Hämäläinen and Vinni (2006) compared machine learning methods for Intelligent Tutoring System. They tackled the problem where educational data sets are so small that machine learning methods cannot be applied directly. They gave general outlines and recommended variations of naïve Bayesian classifiers which are robust.

As a summary, many studies on predicting academic performance have been carried out and looked at various aspects of influencing factors. Some looked at web-based system data logs, conducted surveys eliciting information which are not available from existing data and explored data mining techniques for analyzing student's performance. This study on the other hand, is to mine useful patterns from the archive of student academic records seeking for influencing factors as performance indicators.

DATA MINING

Data mining is the process of automatically discovering useful information in large data repositories. It is an integral part of Knowledge Discovery in Databases (KDD), which is the overall process of converting a series of transformation steps, from data

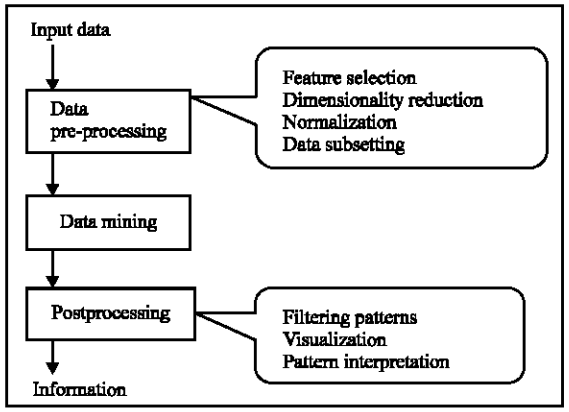

Fig. 1: The process of knowledge discovery in databases (KDD)

preprocessing to post-processing of data mining results. Figure 1 shows the process of knowledge discovery in databases.

The input data can be stored in a variety of formats and may reside in a centralized or distributed data repository. The preprocessing steps transform the raw input data into an appropriate format for subsequent analysis. It is perhaps, the most laborious and time-consuming step. Sometimes, post-processing steps are required to ensure only valid and useful results are incorporated.

Data mining tasks are generally divided into 2 major categories, namely, predictive and descriptive tasks. The objective of predictive tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables. On the other hand, the objective of descriptive tasks is to derive pattern (correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships in data.

Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables. The two types of predictive modeling tasks are classification, which is used for discrete target variables and regression, which is used for continuous target variables. The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable.

The models of decision trees, neural networks based classifications schemes are very much useful in analyzing academic data.

DATA PREPROCESSING

Data is often far from perfect. Often the raw data must be processed in order to make it suitable for analysis. While, one objective may be to improve data quality, other goals focus on modifying the data so that it better fits a specified data mining technique or tool.

Nguyen *et al.* (2007) have conducted a detailed comparison of data mining tools appropriate to predict academic performance. They have chosen the Waikato Environment for Knowledge Analysis (WEKA) in term of computational perspective, wider range of algorithms, better data preparation tools and its support for very large data sets. In this experiment, we only used classifiers provided in WEKA to predict our students' academic performance.

WEKA is a collection of machine learning algorithms and data processing tools. It contains various tools for data pre-processing, classification, regression, clustering, association rules and visualization. There are many learning algorithms implemented in WEKA including Bayesian classifiers, Trees, Rules, Functions, Lazy classifiers and miscellaneous classifiers. The algorithm can be applied directly to a dataset.

In the data preprocessing step, we are particularly interested in analyzing the data for patterns such as which first year courses contribute the most to the overall performance. The real-world data set obtained from the Student Information System does not store sufficient students' background information such as detail admission/entry qualifications, to allow us to perform analysis other than courses taken and the grades obtained. Furthermore, sensitive attributes such as gender and race were not permitted to be used. Some data preprocessing was carried out before the attribute importance ranking was done. For example, a continuous attribute, e.g., CGPA, needed to be transformed into an attribute with discrete categories, e.g., first-class, second class upper, second-class lower, third-class, in order to apply a particular technique.

We collected 2427 complete records of bachelor of computer science students admitted from year 2000 to 2006. We split the data for training (1747 records -72%) and testing (680-28%). Students that failed to complete their studies are not included in our data. There are 396 attributes or subjects registered by the students. Without attributes selection, it makes our data set too large for predicting purpose. Attribute importance analyses were performed in order to rank the attributes by significance in determining the target values as well as to reduce the size of a prediction. Furthermore, it helps to increase speed and accuracy of methods in predicting task. We used the Correlation-based Feature Subset Selection (CfS) and Consistency Subset Selection (CoE) filter algorithm to rank and select the attributes that are most useful for the classifying task. Both of these attribute selection techniques used the BestFirst-D1-N5 searching technique.

The bachelor of computer science students are required to take a total of 102 credits of subjects which comprised of compulsory and elective courses. The compulsory courses are divided into two components, namely university courses (Public Speaking, Management, Malaysian Nationhood, etc.) and core courses (Computer Science, Mathematics). Elective courses can be any courses offered by any faculties.

RESULTS

Table 1 shows the results of the attribute importance analysis using CfS which demonstrates subjects that have strong correlation with graduated students' class. It evaluates the worth of an attribute by measuring the information gain with respect to the class. The number of university courses are 5, 10 elective courses from outside the faculty and the rest 6 courses are computer science and mathematics core courses.

Table 2 shows the results of the attribute importance analysis using CoE which demonstrates subjects that have strong correlation with graduated students' class. It evaluates the worth of an attribute by measuring the information gain with respect to the class. The number of university courses are 9, 8 elective courses from outside the faculty and the rest 7 courses are computer science and mathematics core courses.

Table 3 shows the grouping of data into various categories. The values (A, A-, B+, B, B-, etc.) are actual grades obtained by students. The CGPA for first class is 4.0-3.750, second class upper is 3.749-3.00, second class lower is 2.999 -2.250, third class is 2.249-2.00.

We perform experiment to classify students into 2 classes. The first class is students whose CGPA falls under the first-class and second-class-upper (FirstSecondUpper) category, the second class is students whose CGPA falls under the second-class-lower and third-class (SecondLowerThird) category. We applied different classifiers on our data set and the obtained the results as shown in Table 4.

Table 1: Attribute importance ranking (CfS)

Attribute	Title	Information gain
**SAK3101	Computer programming II	0.166
**SMM3001	Multimedia technology	0.139
**SAK3207	Computer organization and assembly language	0.135
**SAK3103	Discrete structures	0.110
**SAK3100	Computer programming I	0.091
*SKP2101	Malaysian nationhood	0.070
*SKP2202	Asian civilization	0.065
*MGM2111	Organization and business management	0.058
*BBI2410	Skills in grammar	0.031
**MTH3100	Kalkulus	0.029
*KOM3403	Public speaking	0.005
***KOM3101	Introduction to corporate communication	0.005
***KOM2404	Basic interpersonal communication	0.004
***MZK3810	Basic music theory	0.004
***SOS3008	Social communication	0.004
***BBF2401	French language I	0.002
***ANT2001	Society and change	0.001
***BBC2402	Chinese language II	0.001
***KOC3402B	Basic communication strategy	0.001
***KEL2300	Resilient individual development	0.001
***DCE2402	Community development from islamic perspective	0.001

*University courses, **Computer science and mathematics courses, ***Elective courses outside the faculty

Table 2: Attribute Importance Ranking (CoE)

Attribute	Title	Information gain
**SAK3101	Computer programming II	0.166
**SMM3001	Multimedia technology	0.139
**SAK3207	Computer organization and assembly language	0.134
**SAK3103	Discrete structures	0.110
**SAK3100	Computer programming I	0.091
*SKP2101	Malaysian nationhood	0.070
*SKP2202	Asian civilization	0.065
*MGM2111	Organization and business management	0.058
*BBI2410	Skills in grammar	0.030
*SKP2201	Islam civilization	0.027
**SIM3301	Software requirement engineering	0.018
***MTH3002	Introduction to calculus	0.017
***KOM3103	Communication skills in organization	0.007
*BBI2411	Reading and discussion skills	0.007
**SKR3200	Computer communication and network	0.006
***BBI2409	English for academic purposes	0.005
***KOM3101	Introduction to corporate communication	0.005
***EDU3043	Thinking skills	0.003
*BBM2403	Scientific malay language	0.003
***BBS2401	Spanish language I	0.003
***BBC2401	Mandarin language I	0.002
*BBM2404	Malay academic writing	0.002
*BBM2405	Functional malay language	0.0003
***BBI2415	Report writing	0.0001

* University courses, **Computer science and mathematics courses, ***Elective courses outside the faculty

Table 3: Categories of data

Values	Category/Group	Grade value
A-, A	A	4
B-, B, B+	B	3
C-, C, C+	C	2
D, D+	D	1
F	F	0

Table 4: Prediction Accuracy of various classifiers into 2 classes (FirstSecondUpper, SecondLowerThird

		Percentage	
Method	Algorithm	CfS	CoE
Bayes	Naïve Bayes	95.29	95.00
	HNB	95.00	94.27
	AODE	95.29	95.29
	BayesNet	95.15	95.00
Tree	NBTree	94.85	95.15
	REPTree	94.85	94.85
	BFTree	94.26	94.56
	J48	94.71	94.56
Function	RBFNetwork	95.29	94.41
	SMO	95.15	94.71
	Logistic	94.85	94.12

Code	Title	Credits	Code	Title	Credits
SAK3100	Computer programming I	3+1	SAK3101	Computer programming II	3+1
SMM3001	Multimedia technology	3+0	SAK3207	Computer organization and assembly language	3+1
SKP2201	Islam civilization	2+0		Asian civilization	2+0
SKP2101	Malaysian nationhood	3+0	SKP2202	Skills in grammar	3+0
KOM3403^{+}	Public speaking	3+0	BBI2410^{+}	Elective	3
	Total credits	15		Total credits	16

Fig. 2: The first year program structure of the bachelor of computer science program

DISCUSSION

The findings from the attribute importance analyses were very significant. More importantly, the top 9 courses ranked by both CfS and CoE are the same. Four out of the top five courses, namely, Computer Programming II, Multimedia Technology, Computer Organization and Assembly Language and Computer Programming I, are first year core computer science courses. Three of next four courses, namely, Malaysian Nationhood, Asian Civilization and Skills in Grammar are categorized as university courses and are also offered during the first year. This clearly indicates that majority of the first year courses are the influencing courses that contribute to the prediction accuracy of the students' academic performance. Figure 2 shows the first year program structure for the bachelor of computer science program.

A 10-fold cross validation was used to evaluate the prediction accuracy. The prediction results using CfS as the attribute selection technique showed that the Naïve Bayes, AODE and RBFNetwork performed best on our data set with an accuracy of 95.29%. On the other hand, AODE scored the best with an accuracy of 95.29% when using CoE as the attribute selection technique.

Our result is in line with the study by Golding and Donaldson (2006) which stated that first year computer science courses are possible factors determining academic performance.

CONCLUSION

Identifying the attributes that contribute the most significant to the student's academic performance can help to improve the intervention strategies and support services for students who perform poorly in their studies, at an earlier stage. Since, educational data is normally small in size as well as sparse, a lot of effort must be put into the preprocessing steps to ensure the filtering process gives a good model. The objective of this study was to ascertain the major courses that contribute to the overall academic performance within the bachelor of computer science program. The findings of this study displayed a trend suggesting that these courses were a determining factor in predicting performance. This is important as it provides groundwork for further evaluation of the program.

REFERENCES

Chamillard, A.T., 2006. Using student performance predictions in a computer science curriculum. Proceeding of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, June 26-28, Bologna, Italy, pp: 260-264.

Golding, P. and O. Donaldson, 2006. Predicting academic performance. Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference T1D-21, Oct. 28-31, San Diego, CA., pp: 1-6.

Hämäläinen, W. and M. Vinni, 2006. Comparison of machine learning methods for intelligent tutoring systems. Proceedins of the 8th International Conference on Intelligent Tutoring Systems, June 2006, Jhongli, Taiwan, pp: 525-534.

Kotsiantis, S.B., C.J. Pierrakeas and P.E. Pintelas, 2003. Preventing student dropout in distance learning using machine learning techniques. Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Oct. 21, Springer Berlin, Heidelberg, pp: 267-274.

McKenzie, K. and R. Schweitzer, 2001. Who succeeds at University? Factors predicting academic performance in first year Australian university students. Higher Educ. Res. Dev., 20: 21-33.

Merceron, A. and K. Yacef, 2005. Educational data mining: A case study. http://www.it.usyd.edu.au/~kalina/publis/merceron_yacef_aied05.pdf.

Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer and W.F. Punch, 2003. Predicting student performance: An application of data mining methods with an educational web-based system. Proceedings of the 33rd Annual Conference on Frontiers in Education, Nov. 5-8, IEEE Computer Society, Washington, DC, USA., pp: 13-18.

Minaei-Bidgoli, B., G. Kortemeyer and W.F. Punch, 2004. Enhancing online learning performance: An application of data mining method. Proceedings of the 7th IASTED International Conference on Computers and Advanced Technology in Education, August 2004, Kauai, Hawaii, USA., pp: 173-178.

Nguyen, T.N., P. Janecek and P. Haddawy, 2007. A comparative analysis of techniques for predicting academic performance. Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference, Oct. 10-13, Milwaukee, WI., pp: 1-6.

Romero, C., S. Ventura, P.G. Espejo and C. Hervas, 2008. Data mining algorithms to classify students. Proceedings of the 1st International Conference on Educational Data Mining, June 20-21, Montreal, Quebec, Canada, pp: 117-126.

Superby, J.F., J.P. Vandamme and N. Meskens, 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop, (ITS'06), Jhongali, Taiwan, pp: 37-44.

Vandamme, J.P., N. Meskens and J.F. Superby, 2007. Predicting academic performance by data mining methods. Educ. Econ., 15: 405-419.