J Inf Process Syst, Vol.11, No.2, pp.165~172, June 2015
http://dx.doi.org/10.3745/JIPS.04.0015

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Learning to Prevent Inactive Student of Indonesia Open University

Bayu Adhi Tama*

Abstract
The inactive student rate is becoming a major problem in most open universities worldwide. In Indonesia, roughly 36% of students were found to be inactive, in 2005. Data mining had been successfully employed to solve problems in many domains, such as for educational purposes. We are proposing a method for preventing inactive students by mining knowledge from student record systems with several state of the art ensemble methods, such as Bagging, AdaBoost, Random Subspace, Random Forest, and Rotation Forest. The most influential attributes, as well as demographic attributes (marital status and employment), were successfully obtained which were affecting student of being inactive. The complexity and accuracy of classification techniques were also compared and the experimental results show that Rotation Forest, with decision tree as the base-classifier, denotes the best performance compared to other classifiers.

Keywords
Educational Data Mining, Ensemble Techniques, Inactive Student, Open University

1. Introduction

Distance education systems, like open universities, usually apply computer-assisted learning media (i.e., e-learning and e-libraries) instead of face-to-face systems, as in conventional universities, in order to support every academic activity. Therefore, students (distance learners) are supposed to study independently. To be successful in an open university, distance learners are required to actively exercise self-discipline and a strong effort to learn. Nowadays, the characteristics of distance learners and their propensity of being inactive is the most fundamental challenging research topic in the educational data mining and distance education.

Many crucial factors that lead to distance learners being inactive have been successfully identified. These factors included variables related to professions, academics, health, family, education systems, and courses selection [1-4]. In Indonesia, Universitas Terbuka (UT) is responsible for administering distance education. UT provides higher education services for Indonesian citizens, regardless of their age, place, and profession, so as to help them further their studies [5]. However, with the existing education systems, which factors actually lead students to fail their studies was difficult to be determined in order to minimize the number of inactive students.

Manuscript received December 9, 2013; accepted March 31, 2014; onlinefirst April 20, 2015.
Corresponding Author: Bayu Adhi Tama (bayu@unsri.ac.id)
* Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Inderalaya 30662, Indonesia (bayu@unsri.ac.id)

Regardless of the distance between students and the university, UT's management should continue to make improvements by monitoring their students' performance. Meanwhile, the effective and efficient monitoring of a student's performance in an open university has turned from traditional analysis to the use of intelligent data analysis, such as a data warehouse and data mining [6]. Data mining includes the analysis of a large amount of data using mathematical, statistical, artificial intelligence, and machine learning techniques to discover hidden patterns (knowledge) that are currently unknown and that are potentially useful in supporting the decision making process [7-10].

In the academic information systems, all data related to academic activities, such as a student's characteristics, are collected and stored. However, inability of university's management to extract, transform and grasp information in the data into valuable knowledge has become a critical obstacle. Data mining could help the university's management discover valuable knowledge from the large amount of data.

To date, and to the best of our knowledge, employing several state of the art ensemble methods by examining and comparing their complexity and accuracy in education field has not been undertaken. As such, this study aims at carrying out an empirical study of state of the art ensemble methods in the field of educational data mining with a real-world data set. Data was obtained from the student record system at UPBJJ UT Palembang. We utilized several machine learning algorithms, such as Bagging [11], AdaBoost [12], Random Subspace [13], Random Forest [14], and Rotation Forest [15], to cope with the problem of detecting inactive students. A well-known data mining tools, WEKA [16], was also used to examine each classifier's performance. Furthermore, the most valuable patterns from the decision tree (DT) [17] are also presented in this paper so as help UT's management improves their policy concerning education management.

We have divided the reminder of this paper into the following sections: a brief review of related work that has been down on educational data mining is provided in Section 2. Section 3 gives a brief review related to this study. The experiment setup, which includes data collection, experiment results, and analysis, is presented in Section 4. Finally, in Section 5 we conclude the paper by giving a summarization of the results.

2. Related Work

In this section some of the studies related to educational data mining and ensemble methods are reviewed.

Regarding education, data mining is a novel technique that enables knowledge discovery and it supports decision-making and recommendations for university management [18]. The application data mining technique in education gives birth to the term educational data mining research area [19]. Some research has been carried out to predict students' performance in open universities.

Predicting a student's dropping out of the Informatics course at the Hellenic Open University was performed by [20]. They utilized and examined six different algorithms and the study argued that Naïve Bayes performed much better in terms of accuracy as compared to other algorithms.

In 2007, Vandamme et al. [21] used and compared the four data mining techniques such as discriminant analysis, neural networks, random forests, and decision tree. The study aimed at predicting students' performance by classifying students into three distinct groups: the 'low-risk' students, who have a high probability of succeeding; the 'medium-risk' students, who may succeed thanks to measures taken by the university; and the 'high-risk' students, who have a high probability of failing (or dropping out).

The latest work concerning data mining predicting students' performance was undertaken by Koutina and Kermanidis [22]. They employed six well-known data mining techniques, which are the most efficient machine learning algorithms, to predict postgraduate students' performance. From their experiment, Naïve Bayes and 1-NN performed the best in terms of predictive accuracy, as compared to other algorithms. A research survey regarding application data mining in education can be found in [10], [23], and [24]. These papers provide comprehensive review of educational data mining by classifying authors who work in this domain based on their disciplines, models used, tasks, and algorithms.

3. Ensemble Methods

Meanwhile, applications of educational data mining with ensemble methods are increasing due to combining a number of classifiers producing more robust and more accurate predictive accuracy.

An ensemble method was first introduced by Breiman [11]. The creation of multiple versions of predictors to get an aggregate predictor called Bagging was initiated. A similar method, called Boosting, which is used for creating a predictive classifier by iteratively learning from a weighted dataset, was proposed by Freund and Schapire [12]. For each learning step, the weighted dataset would be evaluated based on the classifier's performance. The classifier model with the highest performance would be used to predict the class label. Boosting has many variants, such as LogitBoost regression, AdaBoost, etc.

Ho [13] suggested a combining technique called Random Subspace. Creating and constructing a classifier in a random subspace might solve a small sample sized problem when the number of training objects is relatively small as compared with the data dimensionality [25]. In 2001, Breiman [14] invented a new combining approach called Random Forests. This approach produces a set of trees in such a way that each of the trees depends on the values of random vectors, which produces significant classification accuracy.

Rodriguez et al. [15] utilized the iterative learning approach by using Principled Component Analysis (PCA), which is called a Random Forest. This approach generates a model by training a base-classifier on a randomly selected subspace of the input data. It could be demonstrated that it performs much better than several other ensemble methods on some benchmark classification data sets (i.e., UCI data sets). A combination of the Rotation Forest and AdaBoost was suggested by Zhang and Zhang [26]. This latest work yielded ensemble techniques that had lower prediction errors as compared with the Rotation Forest and AdaBoost.

Mostly, ensemble methods were designed to use DT as a base-classifier. However, a neural network (NN) could perform better, as could a base-classifier. In this paper the accuracy of each ensemble method are assessed with two different base-classifiers.

4. Result and Analysis

In this section data collection, classification analysis, and decision tree rules are presented.

4.1 Data Collection

A dataset was obtained from student record systems from 2011 to 2012. A student record system stores students' detailed demographic and academic history data. The UT used information contained in the student record system as input for entire aspects of decision-making, course and program development, and other academic purposes. The final dataset contained 453 records, where 68.65% (311) cases were in 'Class 1 (inactive)' and 31.35% (142) cases were in 'Class 0 (active).' Based on recommendations from an expert, we determined the 10 significant input variables to be as follows: age, gender, marital status, occupation, scholarship, enrolled semester, cumulative GPA, credits, major, and high school major. A description of each input variables is shown in Table 1.

Table 1. Input variable description

No	Feature	Explanation
1	Occupation	Current students' occupation (unoccupied, government employee, private company, entrepreneur)
2	Subject	Course subject (statistic, math, agriculture, fishery, livestock)
3	Scholarship	Grant that support student expense during school (Boolean) (yes/no)
4	Marital	Student marital status (Boolean) (yes/no)
5	Credit	Cumulative credits (numeric)
6	GPA	Cumulative grade point average (numeric)
7	Enrolled semester	Student enrollment semester (odd semester/even semester)
8	Age	Students' age (numeric)
9	Previous study	Students' last study (senior high school/diploma)
10	Gender	Student's gender (male/female)

4.2 Classification Analysis

We conducted an empirical study using ensemble methods with both DT (also known as J48) and NN as base-classifiers and then used a k-cross validation technique (with k=10) as the performance metric. The evaluation metrics is considered as follows: true positive (TP) is the number of inactive students correctly classified as being inactive students. False positive (FP) is the number of active students incorrectly classified as being inactive students. True negative (TN) is the number of active students correctly classified as being active students. False negative (FN) is the number of inactive students incorrectly classified as being active students. The parameters to analyze the performance of classifiers were measured as follows:

- Recall: Defines the proportion of inactive students correctly classified as inactive students (Recall = TP/(TP+FN)). It is also called the TP rate.
- Precision: Specifies the percentage of student records classified as inactive students. It is also called the positive predictive value (Precision = TP/(TP+FP)).

Table 2. Detailed accuracy of each classifier (%)

	J48	NN	Random Forest	Bagging J48	Boosting J48	Random Subspace J48	Rotation Forest J48	Bagging NN	Boosting NN	Random Subspace NN	Rotation Forest NN
Recall	0.909	0.894	0.901	0.905	0.905	0.890	0.912	0.892	0.894	0.887	0.898
Precision	0.909	0.893	0.900	0.904	0.904	0.889	0.912	0.891	0.893	0.888	0.897

However, when applying DT and NN as base-classifiers on several ensemble methods, such as Bagging, Boosting, and Random Subspace, their accuracy was slightly worse when compared to an individual classifier. These classifiers generally did not outperform single classifiers, which contradicts the conclusion of several empirical studies on Bagging and Boosting [11,12,27], which have stated that Bagging and Boosting improve the performance of single classifiers by reducing bias and variance. Our experimental study that we carried out on a real data set shows that the performance result of combining techniques might be affected by the small sample sized properties of the base classifier.

Meanwhile, the Random Forest were poor in comparison to other methods opposes to previous study [14]. This suggests that a Random Forest gives results that are competitive with Bagging and Boosting. Surprisingly, the Rotation Forest with DT as a base-classifier had the best performance out of all other ensemble methods with a 91.2% in accuracy. Table 2 provides the results of our experiment on determining the classification accuracy of each algorithm with precisions and recalls value metric.

4.3 Decision Tree Rule

A DT uses information gain in order to choose the candidate attributes for each step while developing a tree. It is one of the data mining algorithms that are widely used, because it has high accuracy rates [28]. The DT successfully generated a total of 37 rules. Ten rules classified samples as being Class 1 and 27 rules classifies samples as being Class 0. The most significant rules for Class 0 and Class 1 are presented as shown below.

- R1: IF occupation = government official AND subject = statistic AND marital status = yes
  THEN Class 0
- R2: IF occupation = government official AND subject = math AND credit <= 100
  THEN Class 0
- R3: IF occupation = unoccupied AND marital status = yes AND subject = math
  THEN Class 0
- R4: IF occupation = unoccupied AND marital status = yes AND subject = statistic
  THEN Class 0
- R5: IF occupation = private company
  THEN Class 1
- R6: IF occupation = entrepreneur
  THEN Class 1

However, of the 6 significant rules that were obtained from the DT, we think that the most influential attribute is occupation. It appears in all rules, and therefore, determines a significant class prediction to

Class 0 (inactive) and Class 1 (active). In addition, most of the active students in the university have decent employment as government officials, employees of a private company, and entrepreneurs. Moreover, other attribute dimensions, such as marital status and study major, are important attributes that could be considered by the UT's management in making the right decisions and developing policies regarding higher education management.

5. Conclusions

In this paper we attempted to perform an empirical study in predicting inactive students in open universities. We applied data mining techniques based on ensemble methods to identify the most determinant predictor in predicting inactive students. Several popular ensemble methods have been examined and compared according to their accuracy and the Rotation Forest has slightly better performance than other classifiers with 91% in accuracy.

Resulted rules show that the student demographic attributes (i.e., occupation and marital status) become the most significant predictor in determining inactive students. From the management perspective, these findings are very useful for the decision maker who has an interest education management research. Having an understanding of the underlying determining factors of inactive student could strengthen the competitive advantage of a university.

This research might have some limitations. There are other directions for further research that should be taken. First, it would be interesting to perform cross-sectional research by comparing the characteristics of an open university with a conventional university. Finally, it would be useful if more data samples could be obtained in the future.

Acknowledgement

This work was supported by Sriwijaya University Research Grant under SATEK Research Program 2012–2013.

References

[1] D. R. Garrison, "Researching dropout in distance education," *Distance Education*, vol. 8, no. 1, pp. 95-101, 1987.

[2] D. Kember, T. Lai, D. Murphy, I. Siaw, and K. S. Yuen, "Student progress in distance education: identification of explanatory constructs," *British Journal of Educational Psychology*, vol. 62, no. 3, pp. 285-298, 1992.

[3] N. Shin and J. Kim, "An exploration of learner progress and drop-out in Korea National Open University," *Distance Education*, vol. 20, no. 1, pp. 81-95, 1999.

[4] M. Xenos, C. Pierrakeas, and P. Pintelas, "A survey on student dropout rates and dropout causes concerning the students in the Course of Informatics of the Hellenic Open University," *Computers & Education*, vol. 39, no. 4, pp. 361-377, 2002.

[5] Indonesia Open University, "UT in Brief," 2015; http://www.ut.ac.id/en/ut-in-brief.html.

[6] E. N. Ogor, "Student academic performance monitoring and evaluation using data mining techniques," in

Proceedings of the Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007), Morelos, Mexico, 2007, pp. 354-359.

[7] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: an overview," *AI Magazine*, vol. 13, no. 3, pp. 57-70, 1992.

[8] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. Amsterdam: Morgan Kaufmann, 2006.

[9] E. Turban, R. Sharda, D. Delen, and T. Efraim, *Decision Support and Business Intelligence Systems*, 8th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2007

[10] A. Pena-Ayala, "Educational data mining: a survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432-1462, 2014.

[11] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[13] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, no. 5-32, 2001.

[15] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619-1630, 2006.

[16] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[17] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kauffman, 1993.

[18] C. Vialardi, J. Bravo, L. Shafti, and A. Ortigosa, "Recommendation in higher education using data mining techniques," in *Proceedings of the International Conference on Educational Data Mining*, Cordoba, Spain, 2009, pp. 190-199.

[19] A. Anjewierden, B. Kolloffel, and C. Hulshof, "Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes," in *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007)*, Crete, Creece, 2007, pp. 23-32.

[20] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2003)*, Oxford, UK, 2003 pp. 267-274.

[21] J. P. Vandamme, N. Meskens, and J. F. Superby, "Predicting academic performance by data mining methods," *Education Economics*, vol. 15, no. 4, pp. 405-419, 2007.

[22] M. Koutina and K. L. Kermanidis, "Predicting postgraduate students' performance using machine learning techniques," in *Proceedings of the 7th IFIP WG 12.5 International Conference on Artificial Intelligence Applications and Innovations (AIAI2011)*, Corfu, Greece, 2001, pp. 159-168.

[23] C. Romero and S. Ventura, "Educational data mining: a survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135-146, 2007.

[24] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601-618, 2010.

[25] M. Skurichina and R. P. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 121-135, 2002.

[26] C. X. Zhang and J. S. Zhang, "RotBoost: a technique for combining Rotation Forest and AdaBoost," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1524-1536, 2008.

[27] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on Machine Learning (ICML1996)*, Bari, Italy, 1996, pp. 148-156.

[28] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

Bayu Adhi Tama http://orcid.org/0000-0002-1821-6438

He received bachelor degree in Electrical Engineering from Sriwijaya University, master degree in Information Technology from University of Indonesia in 2004 and 2008, respectively. Since March 2015, he is with the Lab of Information Security and Internet Applications, Department of IT Convergence and Application Engineering, Pukyong National University (PKNU) as a Ph.D. candidate.