Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100

Erkan Er

***Abstract*—In this paper, a model for predicting students' performance levels is proposed which employs three machine learning algorithms: instance-based learning Classifier, Decision Tree and Naïve Bayes. In addition, three decision schemes were used to combine results of the machine learning techniques in different ways to investigate if better classification performance could be achieved. The experiment consists of two phases that are testing and training. These phases are conducted at three steps which correspond to different stages in the semester. At each step the number of attributes in the dataset has been increased and all attributes were included at final stage. The important characteristic of the dataset was that it only contains time-varying attributes rather than time-invariant attributes such as gender or age. This type of dataset has helped to learn to what extend time-invariant data has significant effect on prediction accuracy. The experiment results were evaluated in terms of overall accuracy, sensitivity and precision. Results are discussed compared to results reported in the relevant literature.**

***Index Terms*—Machine learning, online learning, students' performance prediction.**

I. Introduction

There are many studies in e-learning field that investigated the ways of applying machine learning techniques for various educational purposes. One of the focuses of these studies is the predicting dropout rates or at-risk students in distance courses by majorly examining log data obtained from learning management systems (LMSs).

The study conducted by Kotsiantis et al [1] is one of the initial studies which investigated application of machine learning techniques in distance learning for dropout prediction. In this study, time-invariant and time-varying data were included and totally six machine learning techniques was employed, which are Decision Trees, Neural Networks, Naïve Bayes algorithm, Instance-Based Learning Algorithms, Logistic Regression and Support Vector Machines. This study was composed of two experimental stages, training and testing. During these stages, number of attributes was increased step-by-step. For example, while only demographic data was included in the first step, data from the first face-to-face meeting was added in the next step. Six algorithms were tested for each these subsequent steps and then they were compared. The important conclusion of this study is that Naïve Bayes algorithm is very successful in the prediction of dropouts; it predicts with 83% accuracy.

Another important study regarding dropout prediction in e-learning courses was conducted by Lykourentzou, Giannoukos, Nikolopoulos, Mpardis and Loumos [2]. In this study, a dropout prediction method was proposed by combining three popular machine learning techniques, which are feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP. For the combination, three different decision schemes were employed. Both time-invariant and time-varying data were used in training and test phases. This study suggested that a combination of machine learning techniques may result in a better prediction accuracy than a single algorithm. According to results, using decision scheme increased the accuracy in student classification to 97-100%.

Moseley and Mead [3] conducted a related study in which they employed machine learning algorithm to predict drop-outs in nursing courses. In this study, rule induction method was used. To implement that method, CHAID (Chi-Square Automatic Interaction Detector) was employed in a decision tree. Two sorts of data were used in testing and validating phases: time-invariant (e.g. age, gender, etc) and time-varying (e.g. grades, attendance). The success of the proposed system was evaluated based on three factors: sensitivity, specificity and overall accuracy. The results of the study showed that the system was able to identify 84% of students who later withdrew prematurely. Of the students who are identified as at risk, 70% eventually withdrew prematurely.

Using both time-varying and time-invariant data seem to be one of the strengths of these studies. However, these studies showed that using solely time-invariant data such as gender or experience did not result in accurate classification initially and inclusion of time-varying data in next steps has increased the accuracy. Therefore, there is no conclusion about success of classification when time-invariant data was excluded and only time-varying data was used.

Furthermore, classification of data into only two groups is another common weakness. The main purpose of these studies was only detecting at-risk students instead of determining performance levels of students. However, classifying students according to their performances in different levels (e.g., average performance, poor performance, worst performance, etc.) might be more useful. In this way, instructors can provide more adaptive feedback for each student. As a result, limited number of classification seems to be another important weakness of these studies.

The common strength of study 1 and study 2 is that they employed more than one machine learning algorithms to investigate which algorithm produces better classification. Although study 1 employed higher number of machine learning algorithms (i.e. 6), study 2 is more effective in terms of evaluating results of different algorithms. To put it another

Manuscript received May 12, 2012; revised July 19, 2012.

Erkan Er is with the Department of Information Systems, Informatics Institute, Middle East Technical University, Ankara, Turkey (e-mail: er@metu.edu.tr).

way, in study 1 each algorithm was evaluated separate from each other and most effective one was determined, on the other hand, in study 2 some decision schemes were used to combine the results of algorithms to check if more accurate results could be obtained. As expected, study 2 proved that more accurate classification could be obtained if a combination of algorithms is employed instead of evaluating them separately.

In conclusion, the application of machine learning in distance education is mainly concentrated on predicting dropouts in distance courses. For the prediction, a number of machine learning algorithms were employed in combination or no combination and students were classified generally as at-risk or non-risk. In addition, both time-invariant and time-varying data was included in the related studies.

II. Problem Definition and Algorithm

A. Task Definition

The overall goal of the study is to propose a method for accurate prediction of at-risk students in an online course. Specifically, log data of LMS, called METU-Online, were used to identify at-risk students and successful students at various stages during the course, called Information Systems 100. Students are classified into two groups: successors and failures. Those whose overall grade is equal to or above 60 are considered as successor and the rest is considered as failure.

The course log data of last five years have been archived in METU-Online system, however in this study only last two semesters' data will be used. Because, these two semesters have dataset that is composed of the same attributes. This selection was on a random base; that is students from different semesters were selected to constitute the training data set. Building of test data set was similar.

Before the experimentation, the attributes of the available log data was examined and some potentially-irrelevant attributes were eliminated. Followings were determined as final set of attributes: attendance for each week (i.e. 10 weeks), midterm exam, assignment #1-3 and final exam.

All of those attributes have a continuous value except attendance for each week which could be *yes* or *no*. The rest of the attributes have a value in the range of 0-100. "Total number of messages sent to forum" is excluded because discussion forum was not used effectively in IS 100 course. Also, "overall participation score" is excluded because instructors may assign high participation score to help at-risk student to pass the course and this situation caused inconsistencies and low prediction rate at initial trials. "Total number of access to lecture notes" and "total # of hits for each chapter" attributes are also not included for similar reasons. Moreover, attendance information after the tenth week is neglected because some of instructors may complete the lecture earlier and after tenth week there might be no classes. The common characteristic of those attributes is that they are time-varying. No time-invariant data such as gender or age are included in this study. This is because this study investigates whether using only time-varying data is good enough to achieve accurate classification especially during initial phases.

It is important for instructors to recognize students' performance levels as earlier as possible so that they can act in time to help at-risk students. For that purpose a step-based approach was employed during training phase. In this way, the classification success of algorithms was obtained for different phases of a semester. There were three steps in total each of which comprises all attributes used in previous step:

1) 1st step: Attendance information for first four weeks, grade of 1st assignment,
2) 2nd step: Attendance information for first seven weeks, grade of 1st, 2nd assignments, midterm grade
3) 3rd step: Attendance information for first ten weeks, grade of 1st, 2nd and 3rd assignments, final exam grade, midterm grade.

After dataset for training and testing were prepared, the experiment was started. In the experimentation phase, three decision schemes were used in addition to three individual machine learning algorithms. In the following section algorithm and decision scheme details are provided.

B. Algorithm Definition

Three machine learning algorithms were used separately for the classification of students as failure or successor, which are K-Star, Naïve Bayes and C4.5.

K-Star is one of the instance-based classifiers, "which is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function"[1]. K-Star algorithm is different from other instance-based learners because it employs an entropy-based distance function. Naïve bayes is a Bayesian learning algorithm which assumes that attribute values are conditionally independent, given the classification of the instance. Naïve bayes algorithm builds the hypothesis by counting the frequency of various data combinations within the training examples [4].

C4.5 is developed as an extension to ID3 algorithm and it builds decision trees using information entropy concept similar to ID3.

In addition to these three algorithms, three different decision schemes were employed to combine the results obtained in three machine learning algorithms. To be more specific, in this approach, instead of employing one machine learning algorithm, a combination of these algorithms were employed to achieve more accurate classification.

The implementation was completed in two phases: training and testing. Each machine learning algorithm firstly was applied over training data. Then, testing phase took place, in which actual predictions were obtained. Available log data contains 625 instances. 2/3 of log data (i.e. 450 instances) was reserved for training and 1/3 of log data (i.e. 225 instances) was reserved for testing.

According to results of the study conducted by Lykourentzou, Giannoukos, Nikolopoulos, Mpardis and Loumos [2] combining results of different machine learning algorithms may produce better classification than a single technique. Therefore, three decision schemas similar to ones used in [2] were used to combine results of three machine learning algorithms. Scheme 1 means that if at least one of the algorithms classifies student as a failure than this student will be considered as failure otherwise successor. Scheme 2

[1] http://wiki.pentaho.com/display/DATAMINING/KStar

means that if at least two algorithms classify student as a failure than this student will be considered as failure otherwise successor. Finally, scheme 3 means that if all three algorithms classify student as a failure than this student will be considered as failure otherwise successor.

Thus, in addition to results obtained from implementation of each single machine learning technique, results of these three decision schemes were considered in evaluation. Therefore, there were 6 distinct result sets at each step. In this way, the accuracy of each technique at each stage was evaluated and the one that produces better classification was identified.

III. Experimental Evaluation

The evaluation of performance of each machine learning techniques and decision schemes were in three folds: overall accuracy, sensitivity, precision.

The overall accuracy was used to measure the proportion of the students whose performance level is correctly predicted by each technique. That is, overall accuracy is equal to percentage of correctly identified students' statuses (e.g. success or fail) in the whole population. It is calculated by the number of correctly identified successors plus number of correctly identified failures divided by total number of students.

Sensitivity criterion was used to measure efficiency of each technique in correctly identifying students' final statuses, fail or success. It measures the (1) proportion of students whose final statuses are correctly identified as failure versus total number of actual failures, and (2) proportion of students whose final statuses are correctly identified as success versus total number of actual successors.

Precision was used to determine firstly the proportion of the students who were correctly predicted as successor among all successors including incorrect classification of successors, and secondly the proportion of the students who were correctly predicted as failure among all failures that comprise incorrect classifications of failures.

In summary, three performance criteria were employed to evaluate performance of each machine learning technique and decision scheme. At each step, performance measurement was performed for comparison.

IV. Resuts

In this section, results of the experiment are presented considering performances of both machine learning algorithms and decision schemes. Three evaluation criteria have been applied to evaluate the performance of each technique in predicting students' success or failure statuses: accuracy, sensitivity and precision.

A. Overall Accuracy

In the following figure, the overall accuracy of three machine learning algorithms and decision schemes are depicted, where horizontal axis represents the each step of testing and vertical axis represents the percentage of correct predictions by each technique.

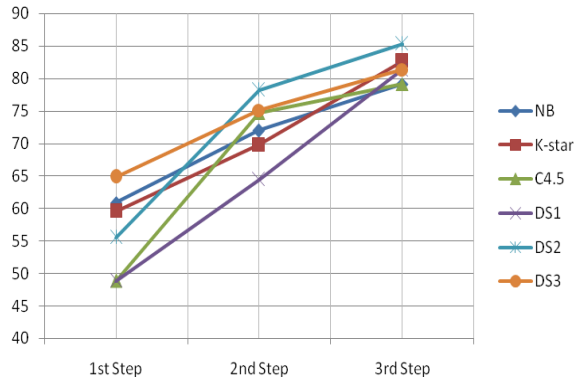

Fig. 1. Overall accuracy of results.

Results showed that decision scheme 3 achieved the best performance with approximately 65 % prediction rate in the first step of testing which comprises only four attributes. It is a significant finding that considerably high accuracy could be achieved in initial stage by using appropriate scheme without including any time-invariant attributes such as demographic characteristics. Naïve Bayes and K-star also performed well in the first step with the accuracy of 60 %. Remaining techniques seem to be inaccurate in the initial phase. Thus, combining multiple techniques using decision scheme approach turns out to be more effective than using a single machine learning technique in the initial phase.

In the second step of testing which contains 10 attributes, each technique resulted in higher accuracy compared to first test. Decision scheme 3 again performed well with the accuracy of 75 %, however decision scheme 2 is found to be superior with the accuracy of 78 %. In this phase, decision schema approach resulted in higher performance than a single machine learning technique, which is similar to findings obtained in the first step.

Accuracy results of each technique seemed to get closer to each other and range of accuracy scores got narrowed to 79-85 % in the final phase. Decision scheme 1 has achieved the highest accuracy in the final stage with the percentage of 85; 93 failures and 90 successors are indentified correctly over 110 and 115 respectively. Following the decision scheme 1, K-star has reached 82 % rate, which is the highest accuracy among the machine learning algorithms. The remaining machine learning algorithms achieved 79% and 81% accuracy. Similar to previous testing phases, in this stage also decision schemes performed better than single machine learning techniques.

Accuracy results indicated that a single machine learning algorithm does not provide accurate estimations over every stage of course and using combination of machine learning, especially decision scheme 2, resulted in higher performance in predicting successors and failures.

B. Overall Sensitivity and Precision

In this section, sensitivity and precision of results are evaluated and analyzed together for both successor and failure predictions separately.

In the Fig. 2 and Fig. 3, overall sensitivity and overall precision of results are depicted respectively, which are calculated by averaging corresponding values (e.g., separate accuracy and precision values) for failures and successors.

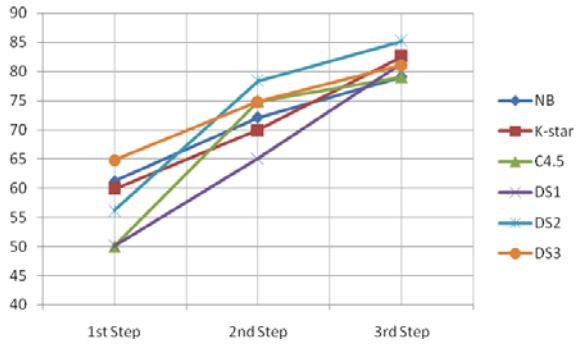

Fig. 2. Overall sensitivity of results.

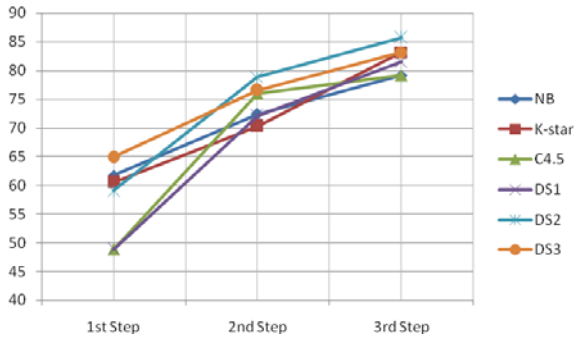

Fig. 3. Overall precision of results.

According to figure 2, sensitivity of results becomes meaningful in second step. Decision scheme 2 has reached to rate of 78% considering the prediction sensitivity of failures and successors. In the final stage, decision scheme again produced the best result (i.e. 85%) in terms of sensitivity. K-Star algorithm and decision scheme 3 also resulted in good sensitivity rates. Overall precision results, which are depicted in Fig. 3, shows a consistency with sensitivity results.

V. Discussion

In this section, key findings of this study are discussed. To begin with, time-invariant attributes were excluded from the dataset to see whether accurate results are obtained in the initial stages with only several time-varying attributes. According to evaluation results, approximately 65% accuracy was achieved in the first step by using decision scheme 3. This result is a little higher than the accuracy obtained in earlier stages in Kotsiantis's [1] study in which time-invariant attributes are also included, and a little lower than the accuracy achieved in Lykourentzou's [2] study which comprises time-invariant attributes again. Therefore, this study showed that exclusion of this type of attributes may ease the process of preparation of dataset and then processing of it during training and testing phases without any negative effect on prediction accuracy. Instead of demographic characteristics of students, using initial attendance and homework grades produces better prediction rate at earlier stages.

Three individual machine learning algorithms were tested using data from METU-Online to determine their overall accuracy, sensitivity and precision. Additionally, three different decision schemes were used to combine the results of machine learning algorithms. The experimental results showed that highest accuracy (82%) was achieved by K-Star among single machine learning algorithms. However, decision scheme 2 has reached to higher accuracy rate, 85.33. Similarly, higher rates are achieved in precision and sensitivity by decision schemes compared to individual machine learning algorithms. K-Star is found to be the most successful machine learning algorithm in terms of accuracy, precision and sensitivity among three individual machine learning algorithm. This shows that instance-based learning algorithms may be more effective in this problem domain.

Furthermore, in this study both successors and failures are analyzed and evaluated in terms of accuracy, precision and sensitivity. Some algorithms produced better results in the classification of successors rather than the failures. This interesting finding may be meaningful in cases where prediction error is high for failures.

In the related literature, studies generally were conducted using both time-invariant and time-varying data in training and testing phases. However, these studies showed that using solely time-invariant data such as gender or experience did not result in accurate classification initially and inclusion of time-varying data in next steps has increased the accuracy. In contrast, this study was conducted by using only time-varying data, which decreased the data organization and computation costs. According to evaluation results, approximately 65% accuracy was achieved in the first step by using decision scheme 3. This result is very close to results obtained in Kotsiantis's [1] study and Lykourentzou's [2] study.

Furthermore, different from the existing studies, this study has employed an instance-based learning algorithm which is K-Star. Results showed that instance-based learning algorithm has achieved more accurate results compared to other individual algorithms, C4.5 and naïve bayes, which were found to be effective machine learning algorithms in that domain.

VI. Conclusion

In this paper, a model for predicting students' performance levels is proposed which employs three machine learning algorithms: instance-based learning classifier, decision tree and naïve bayes. In addition, three decision schemes were used to combine results of the machine learning techniques in different ways to investigate if better classification performance is achieved. The experiment results were evaluated according the overall accuracy, sensitivity and precision and results of the study are discussed compared to results reported in the relevant literature.

Time-invariant attributes are excluded and only time-varying data are used in the proposed study. Experimental results showed that exclusion of time-invariant data has no significant impact on overall results. In other words, using only time-varying data is enough to obtain accurate classification. In literature there is no study that excludes time-invariant data. In this study, the effect of time-invariant data is measured.

Potential future study could be conducting a similar experiment using solely instance-based learning algorithms such as k-star or k-NN. This study has shown that instance-based algorithm, K-star, has produced better results compared to other algorithms. Conducting this study using only instance-based learning algorithms and combining

results of these algorithms again by using decision schemes may result in more accurate results. This could be considered as an important future study.

REFERENCES

[1] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques," *AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, pp. 3-5, September 2003.

[2] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, 2009. "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 950-965, November 2009.

[3] L. G. Moseley, D. M. Mead, "Predicting who will drop out of nursing courses: a machine learning exercise," *Nurse Education Today*, vol. 28, no. 4, pp. 469-475. May 2008

[4] T. M. Mitchell, *Machine Learning*, New York: McGraw-Hill, 1997.

Erkan ER received Master Degree in Information Systems Department from the Informatics Institute, Middle East Technical University, Turkey. He received his Bachelor Degree in Computer Education and Instructional Technologies Department, Turkey. His research interests are machine learning in education, educational data mining, and e-learning.