Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method

http://dx.doi.org/10.3991/ijet.v10i1.4189

Mingjie Tan[1,2], Peiji Shao[1]
[1] University of Electronic Science and Technology of China, Chengdu, P. R. China
[2] Sichuan Open University, Chengdu, P. R. China

***Abstract*—The dropout high rate is a serious problem in E-learning programs. Thus, it is a concern of education administrators and researchers. Predicting the dropout potential of students is a workable solution for preventing dropouts. Based on the analysis of related literature, this study selected students' personal characteristics and academic performance as input attributions. Prediction models were developed using Artificial Neural Network (ANN), Decision Tree (DT) and Bayesian Networks (BNs). A large sample of 62,375 students was utilized in the procedures of model training and testing. The results of each model were presented in a confusion matrix and were analyzed by calculating the rates of accuracy, precision, recall, and F-measure. The results suggested all of the three machine learning methods were effective for student dropout prediction, but DT presented a better performance. Finally, some suggestions were made for future research.**

***Index Terms*—Student Dropout, E-Learning, Prediction, Machine Learning**

I. INTRODUCTION

The scale of E-learning has expanded continuously in the past 10 years due to its unique characteristic of being unconstrained by time or geographical limits. The number of registered students in American colleges and universities who participated in at least one online course from 2002 to 2010 has maintained an annual growth rate of about 10-20%, and in 2010 the number reached 6.14 million, accounting for 31.3% of all registered students [1]. According to statistics from the Chinese Ministry of Education, in 2011, the scale of distance education for bachelor/college students reached 4.53 million persons [2]. Along with the rapid growth of E-learning, its problem of having a much higher student dropout rate than traditional learning has also become more prominent. Studies assert that the dropout rate for E-learning is 10-20% higher than traditional learning [3], while other literature indicates an even higher dropout rate. For example, the dropout rate for the Open University (UK) is as high as 78% [4]. In China, the dropout rate for traditional learning is about 5%, while the dropout rate for E-learning is as high as 15-40% [5-7].

High dropout rates have negative effects on both the educational institutions and students and are not conducive for the healthy development of E-learning. Dropouts increase the average cost per student for education institutions [8], as the cost for recruiting a new student is usually several times that of retaining a potential dropout [4]. From the perspective of students, termination of learning is a waste of their initial economic investment and effort, while the universal phenomena of dropping out is not conducive to the popularization of online learning [9]. In addition, high dropout rates will inevitably lead to lower graduation rates, which may have a negative impact on the social reputation of educational institutions, and in turn, may result in reduced government funding and subsequently lead to a vicious cycle [10]. The United States, Australia, Britain and South Africa all consider student retention rate as an indicator of governmental assessment of the quality of higher education institutions [11].

The means through which student dropout rates can be effectively reduced has become an unavoidable issue in the development process of E-learning, which has received the utmost attention from educational institutions and researchers. Most of the existing empirical research investigates the patterns and reasons for student dropout from statistical patterns of attributions, such as demographic characteristics, semesters lost, course passing rates, and the field of study. Based on empirical analysis, researchers have proposed a series of models to explain the factors for losing online learners and attempting to reduce the loss rate by preventing negative factors while improving positive factors at the macro level. However, as the individual differences of learners are large, improvement strategies on the macro level are often ineffective due to their lack of specificity.

The premise for reducing dropout rates is to understand the various factors associated with dropping out. The key to reducing dropout rates is to make use of these factors to screen out potential dropout students and take targeted retention measures before the dropout behavior happens. This study makes use of the machine learning method for constructing prediction models and uses data from information systems of online education institutions to train the models. After training, the obtained prediction model samples can then be used for predicting dropout behavior that has yet to occur. Online education institutions can make use of this method to identify potential dropouts in a timely manner and to take retention measures before the dropout behavior happens to reduce the dropout rate.

Subsequent sections of this paper include the following: first, selection of input attributions for the prediction model is performed based on a literature review; second, an introduction is provided on the methods used in this study; third, prediction results are presented and analyzed; and

lastly, the research is summarized and possible future research directions are proposed.

II. LITERATURE REVIEW AND ATTRIBUTION SELECTION

Studies on the reasons for or factors of dropping out have a long history. Although these researches were not directly intended for predicting dropout potential, they do facilitate the selection of attributes for the prediction models in this study. Reviews on existing research results relating to the theoretical framework and empirical analysis can help determine the input variables for constructing the prediction models.

A. Theoretical Framework

Tinto's proposal of a higher education dropout model is an early theoretical framework in the field of dropout research [12]. The model proposed that a student's family background, individual factors and previous education are prerequisite factors for student dropout. Whether or not student will drop out depends on the interaction between the student and the learning environment during the learning process. These interactions primarily include academic integration associated with academic performance and intellectual development as well as social integration associated with peer interaction and student- teacher interaction. Tinto's model, although not intended for E-learning, has a relatively important guiding significance for subsequent studies [13]. Based on Tinto's research results, Kember introduced cost-benefit analysis for explaining the decision-making process of students dropping out, emphasizing that variables would change dynamically over time and that there would be multiple occasions in the students' learning process when they decide whether or not to terminate or continue their studies [14].

Bean and Metzner proposed a conceptual model for explaining the reason for non-traditional students dropping out, including the students of distance education [15]. The model stated that students' intention for terminating learning is directly related to their academic performance (scores), academic variables (study habits and academic suggestions, etc.), psychological outcomes (efficacy, satisfaction and pressure, etc.) and environmental variables (economic status, working hours and external encouragement, etc.), whereas prerequisite background variables such as age, gender, registration status, place of residence, educational objectives and previous education can either be directly or indirectly related to the dropout behavior.

Based on a summary of the previous framework and related research, Rovai proposed a comprehensive student retention model that is suitable for E-learning (Figure 1) [16]. The model summarizes the influence factors into 1) students' personal characteristics (age, ethnicity, gender and academic preparation, etc.) and their learning skills (computer literacy, information literacy, time management skills and writing and reading abilities, etc.) prior to enrollment and 2) internal factors (academic integration, social integration, identity with the school and scoring performance, etc.) and external factors (economic situation, working hours, family support and life crisis, etc.) post enrollment. External variables, including individual student characteristics, learning skills and external factors, jointly influence students' decisions on whether or not to keep learning through internal factors. Based on Rovai's

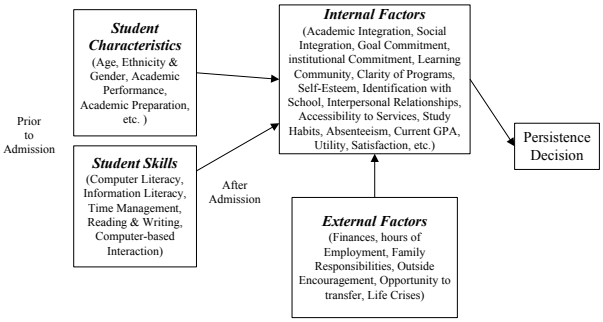

Figure 1. Rovai (2003) model for integrated student retention

model, Park highlighted that external factors influence internal factors both before and after enrollment; students' personal characteristics influence external factors, while a mutual influence exists between the internal and external factors. Based on the above analysis, Park proposed a theoretical framework for dropping out that was applicable to adult E-learning [17]. Rovai's and Park's theories are relatively comprehensive summaries of influence factors of student dropouts supported by subsequent empirical studies [13, 18].

B. Empirical Research

Empirical research in the field of dropping out usually uses the classic dropout theoretical framework as its basis, and makes use of the following four methods to study the factors affecting student dropout: 1) acquire data through questionnaires and analyze the correlation among variables; 2) identify the distribution trends of demographics and other attributions of dropout rate, using the data acquired through questionnaires or obtained from the information system; 3) identify the reasons for dropping out through the analysis of interview transcripts, using the text analysis method; and 4) a combination of the above three methods [19].

Logistic regression is a method frequently used for analyzing the correlation between attributions in studies related to student dropout rates. Roblyer et al. used a Likert scale that contained 60 measurable variables for conducting a survey of 2,162 E-learning students. Binary logistic regression analysis of the sample data revealed a relatively strong correlation between demographic attributions (such as age) and learning achievement (such as GPA) and student dropout [20]. Nichols et al. made use of ordinal regression for analyzing the data obtained from 187questionnaires returned by distance education students; the results showed that among four sets of attributions (namely computer usage attitudes, learning motivation, perception of satisfaction and previous academic performance), only previous academic performance exhibited a strong correlation with dropout [21]. Research by Zhang et al. used multivariate regression to analyze the data of 57,549 students from nine universities: results showed that demographic attributions such as gender, ethnicity and nationality attribution are significantly correlated to dropping out [22]. Another logistic regression study conducted by Doherty analyzed the data of 10,466 students in the educational institution information system; results showed that the students' demographic attributions, learning methods, and curriculum interaction were correlated to dropping out [23].

Other methods have also been used for identifying factors associated with student dropout. Mendez et al. made

use of the Decision Tree (DT) and random forest algorithms for analyzing the data of 2,232 science and engineering students and concluded that attributions such as the academic performance prior to enrollment and in the first year following admission and the number of courses undertaken were related to dropping out [24]. Araque et al. made use of principal component analysis in conjunction with the stepwise regression method to analyze the data of 75,839 students from three universities in Southern Spain; results showed that attributions such as age, parents' education, academic performance, number of courses retaken could be used for explaining dropout behavior [25]. Zhe Ji et al. made use of the structural equation modeling method for proving that social integration, external attribution, academic integration and academic disadvantages had impacts on student dropout [26].

Comparatively, Chinese researchers primarily use descriptive statistics and interviews for investigating the factors associated with student dropout. Ying Li and research team performed statistical analysis on the academic records of 142 student dropouts; results showed that low passing rates were closely related to students' giving up their studies [5]. In other literature, the research team investigated the reasons for student dropping out through a questionnaire (118 students) and interviews (98 students and 40 teachers). The research concluded that in addition to course passing rate, other factors such as work and study conflicts, profession selection error, improper learning methods and emergencies were also key factors influencing student dropout rates [27]. Several Chinese empirical studies found significant differences in dropout rates among students of different genders [28-30], ages [29-31] and study program level, these differences have also been confirmed in some empirical studies abroad.

C. Selection of Attributions

In this study, the attributions used for predicting dropping out, in addition to being correlated to dropping out, also need to be obtainable from the information systems of online education institutions in real-time. Individual student characteristics, which comprise demographic attributions, are information that most educational institutions collect during student registration. Individual student characteristics are set as the prerequisite factor for student dropping out in all of the above mentioned theoretical models, whereby its correlation with dropping out has also been supported by a large amount of empirical research. Students' academic records, which reflect the students' academic performance, are data bound and stored in the information systems of educational institutions. In Tinto and Kember's models, academic performance is reflected as academic integration, which was unrelated to dropping out. However, Bean and Rovai's models proposed that academic performance was directly related to dropping out, which has been validated by a majority of the empirical research.

Related research also identified many other factors associated with dropping out, such as factors involving educational institutions, including the difficulty of the curriculum, academic support services, teacher-student interaction, or study environment factors, (e.g., work study conflict, family support and emergencies) as well as psychological factors (e.g., study motivation, learning orientation and self-efficacy). Despite their association with dropping out, it is difficult or impossible to obtain data related to these factors from the information systems of online education institutions; hence, these factors were not used as input attributions for prediction in this study. However, there is a certain level of correlation between these factors and academic performance and hence can be represented by academic performance to a certain extent.

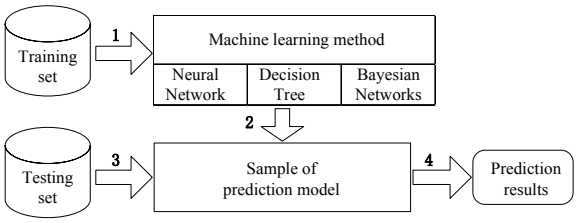

Figure 2. Overall Research Framework

This study used two sets of attributions, namely students' personal characteristics and academic performance, as input attributions in the predictive model. These two sets of attributions were selected based on their close relation to dropout as well as ease of real-time acquisition from the information systems of educational institutions.

III. RESEARCH METHODS

A. Overall Framework

This study makes use of the machine learning method for constructing a prediction model for dropping out to achieve the objective of identifying potential dropouts before the behavior happens. The essence of this study is the construction of a binary classification model in which the samples can be categorized into two classes, namely the dropout class and the retention class. The variables of a single sample are composed of two parts, i.e., the input attribution X = (x1, x2, ..., xn) and the category attribution Y; the process for constructing a classification model is the establishment of a mapping function y = f (X) in which the function can be used to determine the category attribution Y of a sample according to the sample's input attribution X. The overall framework of the study is shown in Figure 2.

The research can be divided into the following four steps: Step 1, Extract attribution data related to student dropouts from the information systems of online educational institutions, construct the training data set and feed the data into the dropout prediction model. Step 2, make use of the data to train the prediction models that were constructed based on machine learning methods such as the Artificial Neural Network (ANN), Decision Tree (DT) and Bayesian Network (BN) to derive the samples of the prediction model. Step 3, extract another section of data from the information systems for constructing a test data set and feed it into the actual samples of prediction model previously generated. Step 4, make use of the prediction model samples to perform predictions on the test data set and evaluate the prediction results generated.

B. Data Description

The source of sample data in this research is from the Open University of China, the largest online educational institution in the country, with a total of 3.59 million enrolled students at present. All courses of the Open University of China are conducted online; the school is an innovative university that provides content-rich courses in

various forms for all members of the community. A relatively high dropout rate has always been a practical issue faced by the school.

The precise definition of a dropout is the foundation of this study, as it determines the principles for categorizing the attributions of sample data categories. Considering the validity of the study period for E-learning students, as stipulated by most higher educational institutions in China, is relatively long (eight years for the Open University of China), students may resume their studies after stopping for a while. Additionally, as very few students voluntarily apply for withdrawal from school, it is difficult to define a real student dropout. Analysis of the school's historical data revealed a strong correlation between students who "did not sit for final exams for two consecutive semesters" and those who "could not graduate within the study period." Table 1 shows that both the confidence level and integrity level are close to 90%. Therefore, this study defines student dropout as one who fails to sit for final exams for two consecutive semesters.

TABLE I.
DATA ANALYSIS FOR THE DEFINITION OF DROPOUT

Grade Enrolled	Number of Students (A)	Number of Non-graduates within the Valid Period (B)	Dropout Rate (B/A)	Number of Students Skipping Final Exam for Two Consecutive Semesters (C)	Total Number of Non-graduates and Students Skipping Final Exam for Two Consecutive Semesters (D)	Confidence Level (D/C)	Integrity Level (D/B)
Spring 2003	12493	2354	18.8%	2225	2148	96.5%	91.2%
Fall 2003	19413	3079	15.9%	3051	2812	92.2%	91.3%
Spring 2004	10645	1793	16.8%	1821	1619	88.9%	90.3%
Fall 2004	12615	1938	15.4%	1978	1663	84.1%	85.8%
Spring 2005	9352	1494	16.0%	1561	1220	78.2%	81.7%
Fall 2005	12866	2229	17.3%	2235	1859	83.2%	83.4%
Total	77384	12887	16.7%	12871	11321	88.0%	87.8%

Sample data in this study are comprised of relevant information from 62,375 students enrolled in the Sichuan Branch of the Open University of China in the three semesters of Fall 2010, Spring 2011 and Fall 2011. Using students who failed to sit for final exams in the Spring and Fall semesters of 2012 as the determination criteria of dropout, the overall dropout rate was 10.4%. The data distribution is shown in Table 2.

TABLE II.
DISTRIBUTION OF RESEARCH SAMPLE DATA

Grade Enrolled	Number of Enrollment	Number of Dropouts	Dropout Rate
Fall 2010	18608	2291	12.3%
Spring 2011	20044	2263	11.3%
Fall 2011	23723	1947	8.2%
Total	62375	6501	10.4%

The data only used the single semester in Fall 2011 as the inspection interval, whereas in practice, we used the end of the semester as the observation point for calculating the dropout rates, which has caused the dropout rates to be lower than what was mentioned in earlier sections of the study. Through the random selection method and in accordance with the data mining practices, we divided the overall data into the training set and the test set in the ratio of 7: 3. The dropout rates of the two data sets were generally consistent after the division.

The attributes of the samples were associated with the students' individual characteristics and academic performance, with a total of 26 variables (Table 3). The majority of the variables in the individual characteristics variables category were obtained directly from the "student information table" in the Academic Management System, while the variables for changes in student age were calculated based on the date-of-birth section. The variables for academic performance were calculated based on student scores obtained from the Academic Management System. Written test results were the scores for the year-end summative exams. Formative assessment results were the scores of homework or tests during the semester, while consolidated results were calculated from the combination of written test results and formative assessment results at a specific ratio.

In order to improve the accuracy of prediction, missing data were filled using the most possible values in the data pre-processing stage, and the continuous numeric variables were discretized based on the requirement of the algorithm. Given the unbalanced characteristics (the ratio between dropouts class and retained class was 1: 9) of the data sets, to ensure that majority class of data was not lost, minority class (dropout class) data were duplicated to balance the data of the training set [32].

C. Machine Learning Techniques

Few studies on the prediction of dropping out in E-learning made use of the machine learning method; however, this method has been relatively widely applied in the telecommunications and financial sectors for customer churn predictions. For example, Tsai et al. made use of the DT and ANN for customer churn predictions for video-on-demand service [33], while Huang et al. used 7 machine learning methods to predict customer churn for mobile communications companies [34]. Research related to customer churn predictions showed that machine learning methods, such as ANN, DT and BN, have relatively good prediction results; therefore, we have also selected these methods for predicting students dropping out in this study. A brief introduction to the three machine learning methods is provided in the following section.

TABLE III.
VARIABLES OF SAMPLE ATTRIBUTIONS

Individual Characteristics				Academic Performance			
Serial Number of Variable	*Name of Variable*	*Meaning of Variable*	*Type of Variable*	*Serial Number of Variable*	*Name of Variable*	*Meaning of Variable*	*Type of Variable*
1	xxdm	study centre	multiset	14	avg_zhcj	average consolidated result	numerical
2	zydm	major	multiset	15	max_zhcj	highest consolidated result	numerical
3	zyccdm	study level	binary	16	min_zhcj	lowest consolidated result	numerical
4	xbdm	gender	binaryus	17	cnt_total	total of studied courses	numerical
5	mzdm	nation	multiset	18	cnt_pass	total of passed courses	numerical
6	zzmmdm	political status	multiset	19	cnt_notpass	total of failed courses	numerical
7	hyzkdm	marriage status	binary	20	pass_rate	pass rate	numerical
8	jgdm	native place	multiset	21	avg_sjcj	average written test results	numerical
9	age	age	numerical	22	max_sjcj	highest written test results	numerical
10	whcddm	educational background before enrollment	multiset	23	min_sjcj	lowest written test results	numerical
11	hkxzdm	live in whether country or town	binary	24	avg_xkcj	average formative assessment result	numerical
12	fbdm	living place	multiset	25	max_xkcj	highest formative assessment result	numerical
13	xflydm	tuition fee source	binary	26	min_xkcj	lowest formative assessment result	numerical

The ANN is composed of input layer units, hidden layer units, output layer units and connections between these layers; it is an algorithms model that simulates the neural networks of animals (Figure 3). The input layer unit corresponds to each variable of the input attributions, while the output layer corresponds to the variables of the category attributions. Training is a process in which the weighting of inter-layer connections is adjusted based on the training using classified data that are already known to achieve a more accurate classification of data with unknown categories. The majority of ANN are based on the multilayer feed-forward error back propagation algorithm [35], which is also the calculation method adopted in this study.

The DT is a tree-structured classification model; the root and internal nodes represent the input values of a certain attribution, the branch represents the output of the input value after the test, and the leaf node represents a specific category (Figure 4). ID3 and CART were the originally proposed DT algorithms, but researchers subsequently proposed the C4.5 and C5.0 algorithms [36], which are improved versions of ID3. All of these algorithms made use of a non-backtracking greedy algorithm. In this study, the C5.0 decision tree classification algorithm was also adopted.

Bayes' theorem is the theoretical basis of the BN, and its essence is a probability network based on probabilistic reasoning. This probability network consists of two parts, namely the directed acyclic graph and the conditional probability table. Each node in the directed acyclic graph represents a random variable, while the conditional probability table is calculated from the data set. The algorithm can be divided into the exact inference algorithm and the approximate reasoning algorithm [37]. To ensure the efficiency of the algorithm, this study adopted the relatively less complex approximate reasoning algorithm.

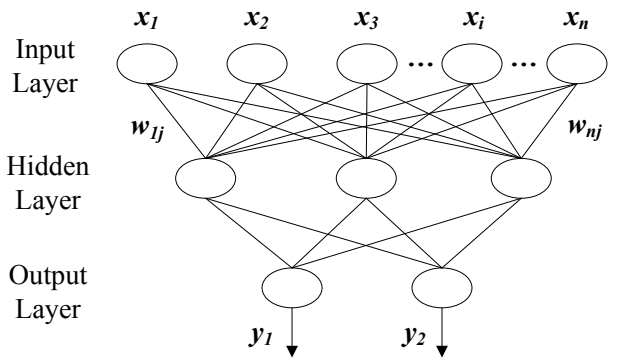

Figure 3. Model of the Artificial Neural Networks Algorithm

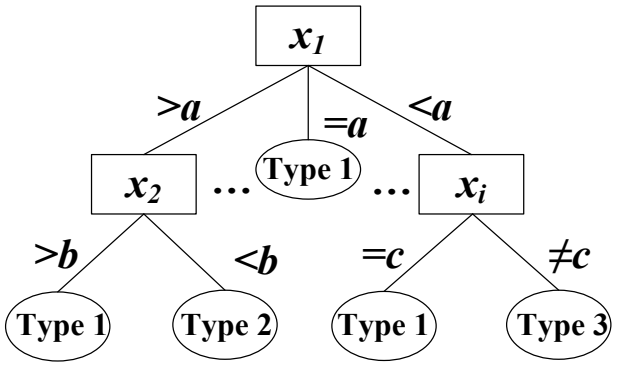

Figure 4. Model of the Decision Tree Algorithm

IV. RESULTS ANALYSIS

A. Methods of Analysis

In this study, we made use of the confusion matrix to present the prediction results of the test, while the effectiveness of the prediction was evaluated using indictors such as the precision rate, the recall rate, the overall accuracy rate, and the F-measure.

If the data set contains n distinct classes, the confusion matrix is an n × n matrix. Given that there were only two types of data (dropouts or retainees) in this study, we have therefore adopted a 2 × 2 confusion matrix (Table 4).

Based on the confusion matrix data, the accuracy rate, the recall rate, the precision rate and the F-measure are defined as follows:

The precision rate of the retained class = $\frac{A}{A+C}$, and the precision rate of the dropout class = $\frac{B}{B+D}$. The recall rate of the retained class = $\frac{A}{A+D}$, and the recall rate of the dropout class = $\frac{B}{B+C}$. The overall accuracy rate = $\frac{A+B}{A+B+C+D}$, and the F-measure = (2 x accuracy rate x recall rate) / (accuracy rate + recall rate).

B. Prediction Results

After dividing the entire data set into the training set and the test set in the ratio of 7:3, the test set contained 18,755 samples. Prediction model samples derived from three different machine learning methods were used for classifying the test data set, and the prediction results are shown in Table 5.

The aforementioned evaluation methods were used for evaluating the results of the prediction (Table 6). In terms of the precision rate and the recall rate of the retained class, there was little difference in the performance of the three prediction models: ANN had the highest precision rate (98.85%), followed by DT (98.33%), and BN had the lowest precision rate (97.60 %). The rankings of the recall rate in descending order are: DT (95.76%), BN (95.67%) and ANN (94.63%). As for the precision rate and recall rate for the dropout class, there are some differences among the three prediction models: the DT had the highest accuracy (63.89%), followed by the BN (63.39%), while the ANN had the lowest precision rate (53.54%). The overall accuracy rate reflects the overall effectiveness of the prediction model. All three models had relatively high overall accuracy rate that exceeded 93%, and the rankings of the overall accuracy rate in descending order are: DT (94.63%), ANN (93.97%) and BN (93.92%).

The objective of this study was to identify potential dropouts. The F-measure value targeted at the dropout class reflects the overall effectiveness of the prediction models in the prediction of the dropout class. The rankings of the models in descending order are: DT (71.91%), BN (69.19%) and ANN (65.65%). In general, the three prediction models were all relatively effective at screening potential student dropouts; comparatively, the DT was the most effective and was more precise at the prediction of the dropout class.

The precision of all three prediction models in the prediction of the retained class was higher than the dropout class, which was mainly because the data attributions were not sufficiently comprehensive. Related research suggests there are many factors affecting student dropping out, and the differences may be large due to individual differences. Since the data in this study were obtained from the academic management systems of online educational institutions, the study is constrained by the comprehensiveness of the data acquired; hence, we have only made use of the students' personal characteristics and academic performance as input variables in the predictive model, which has an impact on the prediction accuracy. Improvements in the machine learning algorithms techniques may also help to increase the precision of the prediction. We have only made use of a single model for prediction in this study, whereas the use of integrated multi-model algorithm will help to improve the precision to some extent.

TABLE IV.
CONFUSION MATRIX

Actual \ Prediction	Retain	Dropout
Retain	A	C
Dropout	D	B

TABLE V.
PREDICTION RESULTS PRESENTED BY THE CONFUSION MATRIX

Predictive / Actual	ANN		DT		BN	
	Retention	***Dropout***	***Retention***	***Dropout***	***Retention***	***Dropout***
Retention	16543	193	16457	279	16335	401
Dropout	938	1081	729	1290	739	1280
Total	17481	1274	17186	1569	17074	1681

TABLE VI.
EVALUATION OF PREDICTION RESULTS

Evaluation Index	ANN	DT	BN
Precision rate of Retained Class	98.85%	98.33%	97.60%
Precision rate of Dropout Class	53.54%	63.89%	63.39%
Recall rate of Retained Class	94.63%	95.76%	95.67%
Recall rate of Dropout Class	84.85%	82.22%	76.15%
Overall Effectiveness	93.97%	94.63%	93.92%
F-measure (of Dropout Class)	65.65%	71.91%	69.19%

V. CONCLUSIONS

Through the summarizations and analysis of relevant literature on student dropout factors in conjunction with the online student attribution data stored in the information system of educational institutions, this study used the personal characteristics and academic performance of students as the input attributions of the predictive model. Three machine learning methods (namely ANN, DT and BN) were used for predicting students' dropout factors. Results showed that all three prediction models were relatively effective at predicting student dropout or retention behaviors; among these, the DT had relatively better prediction results. This study has a certain practical value for resolving the issue of high dropout rates in open and distance education, as it provides online education institutions with a method for screening students with a potential to drop out before the dropout behavior happens.

Using the method proposed in this study, we performed dropout predictions on existing students in the Sichuan Branch of the Open University of China. Subsequently, the list of predicted potential students at risk for dropping out were submitted to the divisions related to students' learning support services so the school can implement targeted measures to retain the potential students before the actual dropout behavior occurs. Actual results proved that the accuracy of the list of the potential dropouts obtained through this method was relatively high.

The ultimate goal for research on student dropout prediction is to improve the precision of the prediction. Given this objective, future research may be conducted from the perspectives of enhancing the attributions and improving the algorithms. First, learning behavior data can be obtained from the academic management system to enhance the input attributions for the predictive model, thereby achieving the goal of improving the precision of predictions; second, improvement can be made to machine learning algorithms, such as the use of integrated model, to improve the precision of the predictions.

REFERENCES

[1] Allen, I. Elaine and Seaman, Jeff. Going the distance: online education in the United States, 2011. Newburyport, USA: Sloan Consortium, 2011.

[2] Ministry of Education, P.R. China. Chinese Education Yearbook Vol. 2012. People's Education Press, 2013.

[3] Doherty, William. An analysis of multiple factors affecting retention in web-based community college courses. Internet and Higher Education, 2006, 9 (4): 245-255. http://dx.doi.org/10.1016/j.iheduc.2006.08.004

[4] Simpson, O.. 22% Can we do better? – the CWP retention literature review final report. http://www.ormondsimpson.com/index.htm (l.v. 2014-1-12). 2011.

[5] Li Ying, Niu Jian, Ding Xia. A Follow-up Study of the Dropouts from the English Program of Open and Distance Learning (part 2) . Open Education Research, 2012, 18 (6): 80-86.

[6] Ran Jingjing, Guo Xuerong. Analysis and countermeasures of dropouts from distance education. Journal of BUPT (Social Science Edition), 2008, 10 (6): 102-106.

[7] Jiang Yulan, Zhou Lei. A research on the dropout of distance learners: Based on Grade 2000 finance majors of NBTVU. Journal of Ningbo Radio & TV University, 2006, 4 (1): 50-56.

[8] Yang Yongjian, Han Xue, Niu Jian, Li Ying. The factors and cost-benefit analysis of student dropout from distance education. China Audiovisual Education, 2011, 7: 64-71.

[9] Chen Lin. Modern distance education high school dropout hazards. Journal of Neimenggu Radio & TV University, 2006,3: 66-67.

[10] Liu Yongquan, Li Ying. Supporting students from success in open and distance learning: An interview with Prof. Ormond Simpson. Open Education Reserach, 2012,18 (5): 4-10.

[11] Blom, K. and Meyers, D. Quality indicators in vocational education and training: International perspectives. Adelaide, Australia: National Centre for Vocational Education Research(NCVER), 2003.

[12] Tinto, V.. Dropouts from higher education: a theoretical synthesis of the recent literature. Review of Educational Research, 1975, 45: 89-125. http://dx.doi.org/10.3102/00346543045001089

[13] Park, J. and Choi H. J.. Factors influencing adult learners' decision to drop out or persist in E-learning. Educational Technology & Society, 2009, 12(4): 207-217.

[14] Kember, D.. A longitudinal-process model of dropout from distance education. The Journal of Higher Education, 1989, 60 (3): 278-301. http://dx.doi.org/10.2307/1982251

[15] Bean, J. P. and Metzner, B. S.. A conceptual model of nontraditional undergraduate student attrition. Review of educational Research, 1985, 55(4): 485-540. http://dx.doi.org/10.3102/00346543055004485

[16] Rovai, A. P.. In search of higher persistence rates in distance education online programs. The Internet and Higher Education, 2003, 6 (1): 1-16. http://dx.doi.org/10.1016/S1096-7516(02)00158-6

[17] Park, J. H.. Factors related to learner dropout in E-learning. Proceedings of International Research Conference in the Americas of the Academy of Human Resource Development. Indianapolis, USA, 2007. 1-8.

[18] Packham G., Jones P., Miller C., et al. E-learning and retention: key factors influencing student withdrawal. Education & Training, 2004, 46(6/7): 335-342. http://dx.doi.org/10.1108/00400910410555240

[19] Lee, Youngju and Choi, Jaeho. A review of online course dropout research: implications for practice and future research. Educational Technology Research and Development, 2011, 59 (5): 593-618. http://dx.doi.org/10.1007/s11423-010-9177-y

[20] Roblyer, M. D., Davis, L., Mills, S. C., et al. Toward practical procedures for predicting and promoting success in virtual school students. The American Journal of Distance Education, 2008, 22(2): 90-109. http://dx.doi.org/10.1080/08923640802039040

[21] Nichols, A. J. and Levy, Y.. Empirical assessment of college student-athletes' persistence in e-learning courses: A case study of a US National Association of Intercollegiate Athletics (NAIA) institution. The Internet and Higher Education, 2009, 12(1): 14-25. http://dx.doi.org/10.1016/j.iheduc.2008.10.003

[22] Zhang, G., Anderson, T. J., Ohland, M. W., et al. Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering Education - Washington-, 2004, 93(4): 313-320. http://dx.doi.org/10.1002/j.2168-9830.2004.tb00820.x

[23] Doherty, W.. An analysis of multiple factors affecting retention in Web-based community college courses. The Internet and Higher Education, 2006, 9(4): 245-255. http://dx.doi.org/10.1016/j.iheduc.2006.08.004

[24] Mendez, G., Buskirk, T. D., Lohr, S., et al. Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. Journal of Engineering Education - Washington-, 2008, 97(1): 57-70.

[25] Araque, F., Roldán, C. and Salguero, A.. Factors influencing university drop out rates. Computers & Education, 2009, 53(3): 563-574. http://dx.doi.org/10.1016/j.compedu.2009.03.013

[26] Ji Zhe, Luo Fafen. Demonstration research on factors affecting learning achievement of distance learners. Open Education Research, 2008, 14 (1): 86-91.

[27] Li Ying, Niu Jian. A Follow-up Study of the Dropouts from the English Program of Open and Distance Learning (part 1). Open Education Research, 2011, 17 (6): 89-97.

[28] Research Group of Jiangxi RTVU. Student attrition in modern open and distance education. Distance Education in China, 2004, 9: 31-36.

[29] Lai Xianming. Statistics and research on online education student attrition rates. Distance Education in China, 2009, 5: 53-57.

[30] Zhang Miaohua, Luo Fafen. Factor analysis and proposal for counter measures of dropouts from modern distance education. Distance Education in China, 2006, 11: 39-44.

[31] Zhu Zulin, Bi Lei, Qi Xinan, etc.. The mining analysis on the dropout rate of modern distance education: Base on the data collected from 1999 to 2009 in Anhui Province of China. Journal of Distance Education, 2011, 4: 18-26.

[32] Cieslak D A, Chawla N V. Learning decision trees for unbalanced data//Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2008: 241-256. http://dx.doi.org/10.1007/978-3-540-87479-9_34

[33] Tsai, C. F. and Chen, M. Y. Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Systems with Applications, 2010, 37(3): 2006-2015. http://dx.doi.org/10.1016/j.eswa.2009.06.076

[34] Huang B, Kechadi M T, Buckley B. Customer churn prediction in telecommunications. Expert Systems with Applications, 2012, 39(1): 1414-1425. http://dx.doi.org/10.1016/j.eswa.2011.08.024

[35] Wong, B. K., Bodnovich, T. A. and Selvi, Y.. Neural network applications in business: A review and analysis of the literature (1988–1995). Decision Support Systems, 1997, 19(4): 301-320. http://dx.doi.org/10.1016/S0167-9236(96)00070-X

[36] Quinlan J R. C4. 5: programs for machine learning. Morgan kaufmann, 1993.

[37] Plant N G, Holland K T. Prediction and assimilation of surf-zone processes using a Bayesian network: Part I: Forward models. Coastal engineering, 2011, 58(1): 119-130. http://dx.doi.org/10.1016/j.coastaleng.2010.09.003

AUTHORS

Mingjie Tan is with the School of Management and Economics at the University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, Sichuan, P.R.China. He is also an associate professor of Computer Science at Sichuan Open University. (e-mail: oliver.tan@163.com).

Peiji Shao is a professor of Information Management at the University of Electronic Science and Technology of China.

Submitted 06 October 2014. Published as resubmitted by the authors 21 February 2015.