Predicting Dropout from Online Education based on Neural Networks

Mingjie Tan[1,2,*] and Peiji Shao[1]

[1]School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, 611731, China; [2]Sichuan Open University, Chengdu, 610072, China

Abstract: While online education keeps expanding, web-based institutions face high dropout rate, pushing costs up and making a negative social impact. Based on the analysis of existing research, personal characteristics and learning behavior were selected as input variables to train a dropout prediction model using neural network algorithm. The outcomes of prediction model were analyzed by calculating the rates of accuracy, precision, and precision. The results suggest this method is effective in identifying potential dropouts, and can help the online education institutions prevent dropout.

Keywords: Dropout, prediction, neural networks, online education.

1. INTRODUCTION

As China pushes harder on its industrial transformation and reconstruction, the society has an increasingly high demand on the quality of employees. Receiving further education has become an important way for people to improve their knowledge structure. Online education has been well received for its openness and flexibility among students who are also in the workforce. According to a statistic report on China's higher education, 1.96 million college and universities students were admitted into 68 web-based colleges (including radio and television universities) around the nation in 2012 [1], taking up 45% of the total number of admitted students into higher educational institutions that year. About 55.7% of people are willing to receive online vocational training, according to statistics from social research institutions. Several large enterprises in China have web-based colleges for staff training.

Online learning market has become mature, but the problem of high dropout rate is a severe challenge that web-based educational institutions face. Student dropout rate for web-based curricula education in China is generally about 20%, much higher than that of traditional classroom education, according to statistics [2]. A high dropout rate is not only a waste of education resources but also a loss for personal education investment. What is more, it leads to a decline in social recognition of e-learning, which is an impediment to the development of web-based education [3]. How to effectively bring down dropout rate has been an urgent problem of e-learning and attracted massive attention from web-based educational institutions.

Existing empirical studies examine the patterns and reasons of learner drain mainly from the statistical perspectives of demographic characteristics, drain terms, exam passing rate, major of study, etc [4-6]. Based on empirical analyses, scholars have explained the factors influencing student dropout of e-learning through many models they proposed, and have been trying to reduce the dropout rate by avoiding negative factors and improving positive ones in a macroscope [7]. However, as there are huge individual differences among students, a macroscopic strategy is often untargeted and ineffective. A feasible measure to reduce dropout rate is to identify the groups of students to quitting so that targeted measures can be taken before attrition happens. This article adopts data mining technology and predicts student dropout based on related attribute data in the learning management system so as to reduce dropout rates in online education.

2. LITERATURE REVIEW

Existing studies on student dropout mainly focus on the various factors that have a significant influence on dropout and build up a macroscopic model to explain the reasons of dropout.

Early normative research about student dropout can be traced back to the dynamic dropout models generation I and II raised by Spady and Tinto [8], which provided explanations for student dropout process. Later, Kennedy and Powell put forward a two-dimensional dropout model to provide explanations based on the influencing factors of student characteristics and environment. Afterwards, Kember proposed a more targeted model for distance education dropout based on the fact that distance education mainly served for students who are also in the workforce [9]. Later, Rovai summarized the factors influencing student dropout from existing studies and proposed an aggregative model to explain e-learning persistence [10].

When Tinto raised his classical dropout model which could provide a good explanation for student dropout in traditional higher education [8] (see Fig. **1**), e-learning was still non-existent. Tinto believes that factors including student family background, individual characteristics and previous education are preconditions of student dropout during learn-

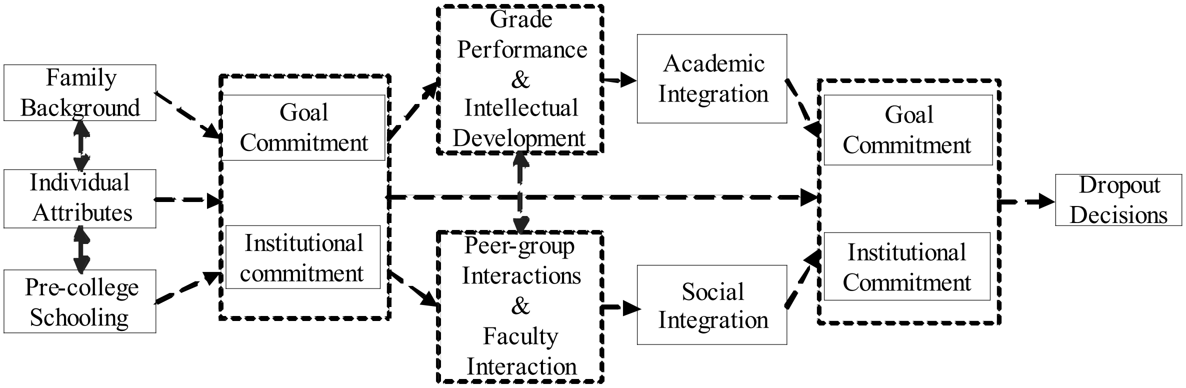

Fig. (1). Dropout model for higher education (Tinto).

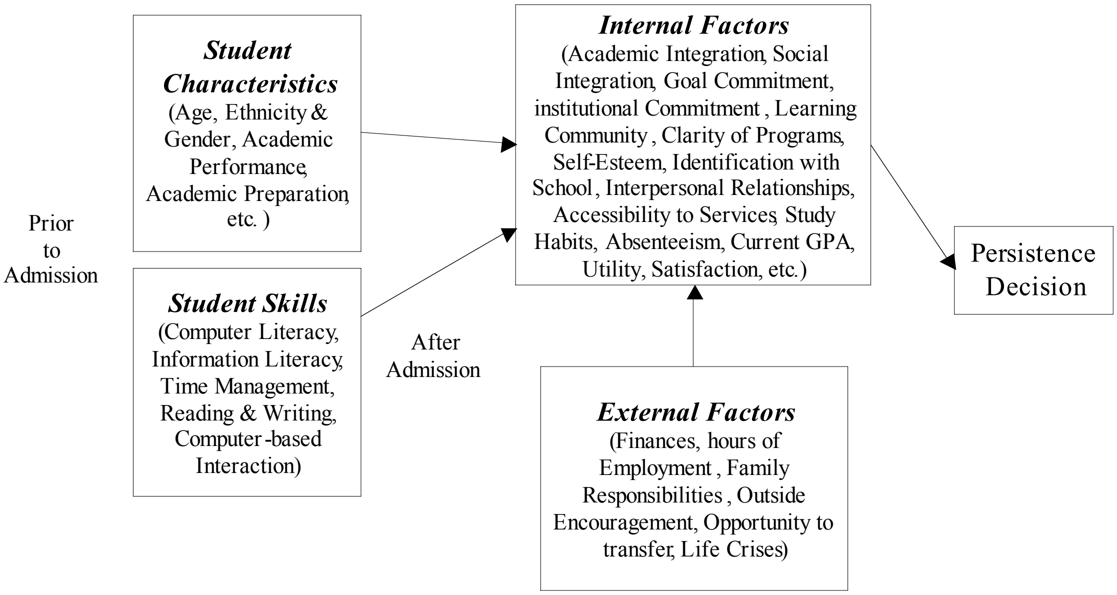
Student Characteristics
(Age, Ethnicity & Gender, Academic Performance, Academic Preparation, etc.)
Student Skills
(Computer Literacy, Information Literacy, Time Management, Reading & Writing, Computer-based Interaction)
Prior to Admission
After Admission
Internal Factors
(Academic Integration, Social Integration, Goal Commitment, institutional Commitment, Learning Community, Clarity of Programs, Self-Esteem, Identification with School, Interpersonal Relationships, Accessibility to Services, Study Habits, Absenteeism, Current GPA, Utility, Satisfaction, etc.)
External Factors
(Finances, hours of Employment, Family Responsibilities, Outside Encouragement, Opportunity to transfer, Life Crises)
Persistence Decision

Fig. (2). Unified model for student persistence (Rovai).

ing. Based on these preconditions, students make commitment with educational institutions to complete their studies in the early stage of learning. During learning processes, students interact with their learning environment at the dimensions of academic integration involving learning results and social integration involving interpersonal relationship. In the process of academic and social integration, students compare the result of continuous fusion with previous promises. If the gap is huge, students may decide to drop out.

Kember's dropout model [9] is mainly proposed for distance education, and its basic ideas are similar to Tinto's model. Based on the fact that distance education mainly serves for students in the workforce, Kember incorporates the impact of students and their social environment on fusion into the model. It also emphasizes on the possible indirect impact of student learning motive on the result of fusion. While learning, students will constantly make cost performance analysis based on the result of fusion, and if the expected learning costs go far beyond benefits, dropout will happen.

The model of Tinto and Kember mainly explains the decision making process of student dropout, which involves some factors related with attrition, but they have not given a comprehensive and systematic summary of various influencing factors. After the above classical dropout models were raised, studies targeted at student dropout came one after another, most of which are concerned about the factors influencing dropout. Based on a systematic classification and summary of existing studies, Rovai proposed an aggregative retention model [10] (see Fig. **2**). Based on this model, external factors, student individual characteristics and learning skills will have an influence on the decision of retention (or dropout) through internal factors.

Although the research results above are not directly linked with the prediction of student dropout, but can serve as evidence for the input attribute selection in a dropout predicting model. Based on the various factors involved in the above models and combining the real situation of information system of online educational institutions, the two groups of attribute data—individual characteristics and academic performances—can be used as the input variables to predict whether students dropout. This is because these attributes are closely related to student dropout and related data can be obtained in real time from the information systems of online educational institutions.

3. METHOD OF PREDICTION

3.1. The Tilt Angle

This paper uses the classification algorithms to set up a binary classification model, which is trained by using related attribute data in the information systems of online educational institutions, so that a prediction model that can divide students into dropout group and non-dropout group. With this model, online education institutions can identify students who are likely to drop out before that happens and take targeted preventive measures to avoid dropout.

Data mining is to extract implicit and useful knowledge or modes from large numbers of fuzzy data in operation information system, of which five basic steps are: attribute selection, data pre-processing, data transformation, model building and outcome assessment (see Fig. **3**).

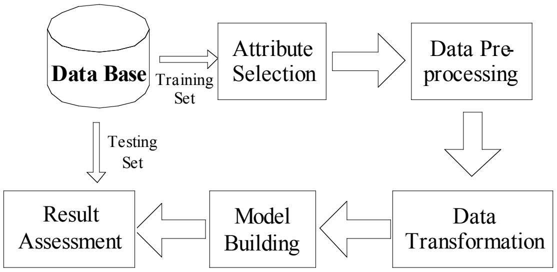

Fig. (3). Basic steps.

The database wherein is that for the learning management systems and education management information systems of online education institutions. Attribute selection is to select attribute data involved with the factors related to student dropout, and it is the starting point of the whole prediction process. A targeted selection of attribute data is the key to increase prediction accuracy. Because there might be loss when information systems collect data, leading to incomplete or conflicting data, pre-processing including data cleaning, integration, transformation and reduction needs to be carried out so as to unify and standardize data. Before data transformation, a proper data mining algorithm needs to be determined. Different algorithms have different requirements for data structure, and data mining is actually transform preprocessed data based on the requirements of selected algorithm. Model building is the core step in data mining. After a model is built using a certain algorithm, the model is then trained with training data set. A trained prediction model can then be used for dropout prediction. Result assessment is to assess the obtained prediction model on its real prediction accuracy with testing data set so as to verify the prediction outcome.

3.2. Classification Algorithm

Classification algorithm can be used to analyze the data in training data set so as to obtain the classification model or rules that describes each category characteristic, and then these model or rules can be used to classify other data sets.

Decision tree, a common classification algorithm, uses its tree structure to represent classification or decision sets, generate rules and discover regulations. Commonly used algorithms of decision tree include the famous ID3 algorithm raised by Quinlan and the later ID4, ID5, C4.5, SLIQ and SPRINT algorithms [11]. Besides, the algorithms below are also used frequently: Bayes classification algorithm—a classification algorithm using probability statistics knowledge [12]. It mainly applies Bayes theorem to calculate the probabilities of each sample belonging to each category, and select the highest one as the ultimate category; neutral network—it uses training sets to continuously train a network of large numbers of nerve cells and then classifies the samples using the trained neutral network; rough set theory—one method that requires no quantitative descriptions of certain characteristics or attributes, but starts from the given problem and find its rules [13]; genetic algorithm—by simulating the biological evolution process, it optimizes and finds solution using three basic operators—selection, recombination and mutation.

This article uses neutral network as the classification algorithm for predicting dropout. This algorithm is connected by the output layer units, hidden layer units and input layer units to simulate the working mode of animal nerves (see Fig. **4**). Each input layer unit corresponds with each variable in the input attribute, and the output layer corresponds with the attribute variables of each category [14]. Through the training of using known category data set in the learning process, the weighted values of connection are adjusted so as to classify the unknown data more correctly. Neutral network model is mostly used for multilayer feed forward neutral networks based on error back propagation algorithm. This paper also adopts this algorithm.

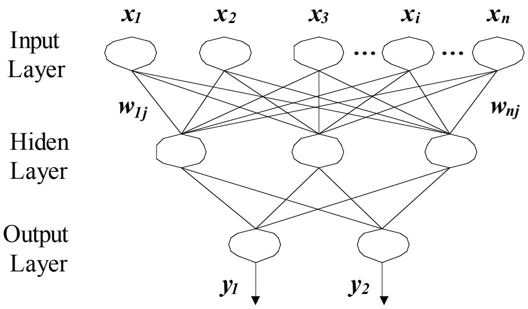

Fig. (4). Neural networks model.

Suppose there are n neurons in the input layer, p neurons in the hidden layer and q neurons in the output layer, and then the operation of the algorithm can be described as follows:

Define the input variable $x = (x_1, x_2, \cdots, x_n)$, input vector in the hidden layer $hi = (hi_1, hi_2, \cdots, hi_p)$, output vector in the hidden layer $ho = (ho_1, ho_2, \cdots, ho_p)$, input vector in the output layer $yi = (yi_1, yi_2, \cdots, yi_q)$, output vector in the output layer $yo = (yo_1, yo_2, \cdots, yo_q)$, expected output vector $d_o = (d_1, d_2, \cdots, d_q)$, the connection weight between input layer and hidden layer W_{ih}, connection weight between hidden layer and output layer W_{ho}, threshold value of hidden neurons b_h, threshold value of output neurons b_o, the number of sample data $k = 1, 2, \cdots, m$, activation function is $f(\cdot)$ and error function $e = \frac{1}{2}\sum_{o=1}^{q}(d_o(k) - yo_o(k))^2$.

The procedures of the algorithm are as follows:

Step 1: Assign a random value between $(-1,1)$ to each connection weight. Set activation function. Set a computational accuracy ε and maximum number of times for learning M.

Step 2: Select the kth input sample $x(k) = (x_1(k), x_2(k), \cdots, x_n(k))$ and its corresponding expected output $d_o(k) = (d_1(k), d_2(k), \cdots, d_q(k))$ randomly.

Step 3: Calculate the input and output neurons in hidden layer.

$$hi_h(k) = \sum_{i=1}^{n} w_{ih}x_i(k) - b_h \quad h = 1,2,\cdots,p$$

$$ho_h(k) = f(hi_h(k)) \quad h = 1,2,\cdots,p$$

Step 4: Use the expected output and real output to calculate the partial derivative $\delta_o(k)$ of error function for output neurons.

Step 5: Use $\delta_o(k)$ and hidden output neurons to correct the connection weight $w_{ho}(k)$.

$$\Delta w_{ho}(k) = -\mu \frac{\partial e}{\partial w_{ho}} = \mu\delta_o(k)ho_h(k)$$

$$w_{ho}^{N+1} = w_{ho}^{N} + \eta\delta_o(k)ho_h(k)$$

Step 6: Use $\delta_h(k)$ of hidden neurons and input neurons to correct the connection weight $w_{ih}(k)$.

$$\Delta w_{ih}(k) = -\mu \frac{\partial e}{\partial w_{ih}} = \delta_h(k)x_i(k)$$

$$w_{ih}^{N+1} = w_{ih}^{N} + \eta\delta_h(k)x_i(k)$$

Step 7: Calculate the global error E.

$$E = \frac{1}{2m}\sum_{k=1}^{m}\sum_{o=1}^{q}(d_o(k) - y_o(k))^2$$

Step 8: Decide whether the error is within tolerance. When the error meets the set accuracy or the number of times for learning reaches the maximum, the algorithm ends. Otherwise, select the next sample and its corresponding expected output. Back to step 3, another round begins.

3.3. Attribute Selection and Data Preprocessing

A large number of studies have shown that students' individual characteristics and academic performances may have a relationship with dropout. Therefore, this article chooses the above two groups of properties as the input variables of the prediction model. Among them, students' personal characteristics include gender, age, location, nationality, etc. Academic performances include the final grade, formative grade, the number of subjects of exam participation, the pass rate, etc.

Based on requirements neural network algorithm has for the types of input variable data, continuous numerical variables such as grade are discredited. Due to the low proportion of dropout group (about 10%), the data belongs to the unbalanced data sets, and therefore the data is balance processed. After processing, erosion and dropout group and non-dropout group take up equal proportions.

4. RESULT ASSESSMENT

In this paper, the data comes from relevant information of 2,354 students enrolled in an online education institution information system in the spring of 2011. If students has not been registered for either of the two seasons—the spring and autumn in 2012—for courses, then they are determined to be dropouts. According to the above decision rule, non-dropouts are 2,103 students, taking up 89.3%; the number of dropout students is 251, accounting for 10.7%. According to the 7:3 proportion of the number of samples, the above data is randomly divided into two parts: the training data set that can be used to build a prediction model and the testing data set that can be used to assess predicting results. After division, the training data set has 1,678 samples, with 1,494 non-dropout samples. The number of dropout samples is 184, with a dropout rate of 11%; Testing data set sample was 676, among which 609 are non-dropouts and 67 are dropouts, with a dropout rate of 10%. Test set and training set have basically the same dropout proportion.

The evaluation criterion of outcome is the confusion matrix shown in Table **1** where A is number of samples who are actual dropouts and are predicted to be dropouts as well; B is the number of students who are dropouts but are predicted as non-dropouts; C is the number of students who are non-dropouts but are predicted as dropouts; D is the number of students who are and are predicted as non-dropouts. Using indicators such as accuracy, hit rate and coverage rate to assess the outcome of the prediction model, the accuracy = (A + D)/ (A + B + C + D), sensitivity= A/ (A + C) and precision = A/ (A + B).

Table 1. Confusion matrix.

The State of the Samples	Predicted as Dropout	Predicted as Non-dropout
Dropout	A	B
Non-Dropout	C	D

After the model was trained using training sets, the prediction model is used to classify the testing data set, the results of which are shown in Table **2**. Calculation according to the above evaluation index shows that accuracy is 89.6%, sensitivity is 48.4% and precision is 67.2%. As accuracy reflects the overall prediction accuracy of the model, so the overall prediction effect is pretty good. As the sensitivity reflects the accuracy of prediction of dropout group, a low hit rate means that the samples predicted as dropouts contain a certain amount of non-dropouts. Precision reflects the probability of this model to identify dropouts, so this prediction model can find 2/3 of potential dropouts.

Table 2. Classification results.

The State of the Samples	Predicted as Dropout	Predicted as Non-dropout
Dropout	45	22
Non-Dropout	48	561

The above analysis shows that the prediction model obtained in this paper can quite accurately identify possible dropouts before dropout behavior happens, although there are some errors. On the whole, it is of value for online education institutions to take measures in advance to prevent student dropout.

CONCLUSION

Student dropout, a real problem faced by the industry of web-based learning, has been studied mostly by exploring the factors influencing student dropout. This paper uses neutral network classification algorithm to predict student dropout through two groups of attributes—student individual characteristics and academic performances in the information systems of online education institutions. This method can help online education institutions to identify potential dropouts before they quit and take preventive measures against it. Assessment of the obtained prediction model proves that the prediction outcome is effective. Further research can enhance the accuracy by enriching attributes and improving the algorithm.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Ministry of Education, P.R. China, *Chinese Education Yearbook*, People's Education Press, China, vol. 2012, 2013.

[2] Y. Li, J. Niu, and X. Ding, “A follow-up study of the dropouts from the english program of open and distance learning,” *Open Education Research*, vol. 18, no. 6, pp. 80-86, 2012.

[3] O. Simpson, “The impact on retention of interventions to support distance learning students,” *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 19, no. 1, pp. 79-95, 2004.

[4] M. Rosli, I. Ismail, R. M. Idrus, and A.A. Ziden “Adoption of mobile learning among distance education students in universiti sains malaysia,” *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 4, no. 2, pp. 24-28, 2010.

[5] Y. Lee and J. Choi, “A review of online course dropout research: implications for practice and future research,” *Educational Technology Research and Development*, vol. 59, no. 5, pp. 593-618, 2011.

[6] J. Park and H. J. Choi, “Factors influencing adult learners' decision to drop out or persist in online learning,” *Educational Technology & Society*, vol. 12, no. 4, pp. 207-217, 2009.

[7] Y. Levy, “Comparing dropouts and persistence in e-learning courses,” *Computers and Education*, vol. 48, no. 2, pp. 185-204, 2007.

[8] V. Tinto, “Dropouts from higher education: a theoretical synthesis of the recent literature,” *Review of Educational Research*, vol. 45, pp. 89-125, 2007.

[9] D. Kember, “A longitudinal-process model of dropout from distance education,” *The Journal of Higher Education*, vol. 60, no. 3, pp. 278-301, 1989.

[10] A. P. Rovai, “In search of higher persistence rates in distance education online programs,” *The Internet and Higher Education*, vol. 6, no. 1, pp. 1-16, 2003.

[11] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Morgan Kaufmann Publishers: Massachusetts, 1993.

[12] I. Rish, “An empirical study of the naive bayes Classifier,” *IJCAI Workshop on Empirical Methods in AI,* 2001.

[13] Z. Pawlak, S. K. M. Wong, and W. Ziarko, “Rough sets: probabilistic versus deterministic approach,” *International Journal of Man-Machine Studies*, vol. 29, no. 1, pp. 81-95, 1988.

[14] B. K. Wong, T. A. Bodnovich, and Y. Selvi, “Neural network applications in business: a review and analysis of the literature (1988-1995),” *Decision Support Systems*, vol. 19, no. 4, pp. 301-320, 1997.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014