Design and Implementation of Early Warning System Based on Educational Big Data

Zhuping Wang, Chenjing Zhu, Zelin Ying, Ying Zhang, Ben Wang*, Xinyu Jin, Huansong Yang
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China

Abstract—**With the continuous popularization of higher education, the academic problems of university students are constantly emerging. Due to the lack of a systematic learning guidance system, students are lack of learning ability, poor binding force and strong dependence. Because of disciplinary violations and academic problems, quite a few students have been delayed in graduation, processed or even dropped out of school. In order to improve this situation as soon as possible, many colleges and universities have established academic warning system one after another. In the previous systems, it is basically based on performance score, credit score and other performance data, and then different warning levels are manually recorded. Without comprehensive relevant data, the single inefficient forms can not guarantee the effectiveness of academic monitoring and early warning. Dependent on the data of teaching and library, this paper suggests an academic early warning system in Hangzhou Normal University. Considering the data of educational administration, library borrowing and self-study, an early-warning model of learning is established after comprehensive analysis. By this model, we can discover and identify the existing and potential academic problems of students in the early stage of college, and inform themselves and their parents to urge students to correct their attitude and study more efficiently.**

Keywords-Academic early warning system; education big data; data mining

I. INTRODUCTION

With the elite education turning to popular education, the academic problems of college students are constantly emerging. Due to the lack of a systematic learning guidance system, students lack learning ability, with poor self-management ability and strong dependence. Quite a few students graduate late or even drop out of school due to violations of discipline and academic problems. The actual demand makes colleges and universities begin to deploy academic early warning system and try it out according to their real situation. "Academic early warning" refers to the information technology in the management of students' academic work, including early detection their potential or real academic problems. It can inform their parents and themselves the possible adverse consequences. At the same time, the corresponding assessment should be taken to prevent college students from leaving normal learning track [1]. It is able to warn students early before the performance is down to irretrievable. Yang B. L. and Guo Z. H. (2012) put forward the early warning of experimental teaching achievement [2]. Zhang H. Y. (2010) [3], Fan B. (2011) [4], Yang B. L. and Guo Z. H. (2012) [2] proposed the need for early warning of students' course selection. Wang Z. H. and Luo B. Q. (2011) formulated a three-level early warning standard according to the students' family background, learning attitude, and academic achievements at the time of their admission [5]. Robert and Joanna argued that attendance, behavior and course performance are the criteria for judging whether a student deviates from the normal track, abbreviated as "ABC criteria" [6].

The research and implementation of above early warning systems are mostly based on the results of data, making an ex post warning, so the management efficiency is low. It only plays a certain role in a certain range, and can not find the students' academic problems in the early stage. In this study, through the integration and analysis of Hangzhou Normal University's education data, we choose the algorithm with higher accuracy to build the model, and design and implement the academic early warning system (AEWS). To identify problems that students have one or two years in advance, the system actively informs students of their current academic performance. At the same time, it prompts and warns students, their parents or teachers, to guide and help students to handle learning problems. As a result, the management of student status is flexible, reducing the adverse student status changes in colleges and universities. The AEWS tries to put parents into the management chain, allow parents to keep track of their children's learning situation at any time, and give full play to the trinity of "students, schools, parents" education function. The data of colleges and universities are very similar, and the system can be applied to other colleges and universities with a little modification. It should be a beneficial attempt to utilize the large data of their education. The AEWS is conducive to comprehensively deepening educational reform and realizing the healthy development of higher education.

II. SYSTEM DESIGN

A. General Design

1) Destination layer

There are four purposes of early warning: to reduce the dropout rate, to promote academic success, to enhance the effectiveness of learning, and to improve the employment rate. The determination of early-warning purpose points out the direction for establishment and operation of early-warning system. It will directly affect the collection and acquisition of data. And it also affect the content layer, mode layer, and result

978-1-7281-0120-0/18/$31.00 ©2018 IEEE

layer. The destination layer is the basis of the whole early warning process.

2) Method layer

The method layer includes data acquisition and pretreatment data analysis. Data acquisition needs to be clear about what technology to collect and what data to collect. Data preprocessing handles data integration, cleaning, and integration. In the era of big data, all-round data of students, with different aspects, are collected for processing and analyzing.

3) Content layer

The content layer solves the early warning problem, including principal component analysis and early warning algorithm screening. Principal component analysis is used to identify the main risk factors from the pre-processed data. And early warning algorithm screening is proposed to determine the optimal early warning model by comparing the prediction accuracy of several classification prediction algorithms.

4) Result layer

The result layer focuses on the presentation of early warning information and intervention strategies. Considering the target level, content level and mode level, this layer determines the presentation of early warning information, and provides intervention strategies. The presentation of early warning information is an intuitive manifestation of AEWS. And intervention strategies include system intervention and manual intervention.

B. Function module and business flow design

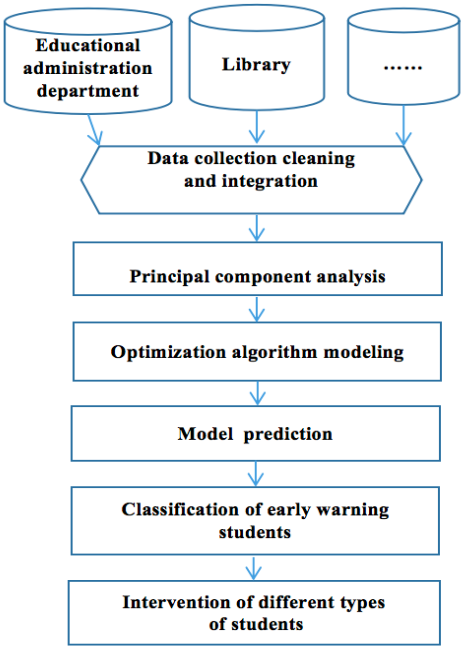

Figure 1. System function module and business flow

Function module and business flow are shown in Figure 1. Firstly, the comprehensive data of students were collected from the educational administration department, library and other departments. After pretreatment, the key risk factors were identified by principal component analysis. Secondly, machine learning classification algorithms (decision tree, Bayes, SVM, neural network, etc.) in data mining are used to train the risk factors and design the optimal academic early warning model. Finally, the model is used to screen out students who may have difficulties in their studies and classify them according to different degrees.

III. DESIGN AND IMPLEMENTATION OF ACADEMIC EARLY WARNING MODEL

In the overall framework of AEWS, prediction modeling is a key link. A case study of Hangzhou Normal University is explored the process of establishing an academic early warning model.

A. Data source and preprocessing

At present, all departments of colleges and universities have realized information management, but the data of departments are independent, forming a number of information islands. In our study, firstly, we need to obtain the data related to students from various departments, such as the student's score sheet, attendance sheet, impoverished student sheet of the student management system, the borrowing information sheet and the credit card record sheet of the library management system, and the entrance guard information sheet of the student dormitory management system. Except that, we also consider the warning of the average credit hour scores and the degree courses, and other data of students in all aspects of study and life. Combining above students' information, the data foundation is built up for academic early warning.

Because of difficulties in the data export of some departments, this study mainly collected the data of the local educational administration system of Hangzhou Institute of Information Engineering, the number of Library borrowings and visits, the warning of the average credit point and degree courses. According to different majors, the students' academic achievement information, library data and warning data from 2009 to 2016 in this institute are integrated, cleaned, and dimensionally reduced.

The data of educational administration system, library data warning data integration, cleaning and dimensionality reduction are shown in Table I (1712 students in total). The average number of Library borrowings and visits is calculated annually. The results are based on the information of average credit points warning, degree course warning and drop-out warning in each semester. On the basis of the above, we further classify and label the warning data, and classify the label with 0-6 numbers (label classification criteria: 0 indicates no warning, 1 as degree warning, 2 as grade point warning, 3 as warning, 4 as drop-out warning, 5 as extension of school system, 6 as graduation). The data table after class processing is shown in Table I. By classifying labels, academic situation of a student is also defined. The larger the tag value means the more warning the student receives, and the worse his academic performance is.

After data processing, the classified data are proper for further analysis. In data mining technologies, this stage is quite important and cannot be omitted.

TABLE I. DATA TABLE AFTER LABEL CLASSIFICATION

School year / Semester											Classified label
No.	Full name	Total scores	Number of courses	Total credits	Credit	Falling number	Rate of passing	Ave. credit hour points	Rank -ing	...	
...	...	665	9	22	19	3	86.36%	2.13	32	...	1
...	...	1092	13	28.5	28.5	0	100.00%	3.35	10	...	0
...	...	1003	14	28	20	8	71.43%	1.86	80	...	2
...	...	1067	13	28.5	28.5	0	100.00%	2.93	21	...	0
...	...	1058	12	26.5	26.5	0	100.00%	3.84	2	...	0
...	...	989	12	27.5	27.5	0	100.00%	3.21	14	...	0
...	...	1021	13	28.5	25.5	3	89.47%	2.56	26	...	0
...	...	721	11	28.5	23.5	5	82.46%	1.45	44	...	4
...	...	...	...	...	...	...	...	...	...	...	...

B. Principal component analysis (identify key factors)

In the study of empirical research, many factors should be taken into account in order to analyze problems comprehensively and systematically. These factors are generally referred to as indicators, and they are also called variables in multivariate statistical analysis. Each variable reflects to varying degrees of information about the problem under study, and there is a certain correlation between the indicators. Therefore, the information from the statistical data has a certain degree of overlap. When using statistical methods to study multivariable problems, too many variables would increase the computational efficiency and the complexity of analytical problems. In the process of quantitative analysis, fewer variables and more information should be analyzed. Principal component analysis is precisely adapted to this requirement and is an ideal tool for solving such problems [7].

Since there is a certain correlation among many variables in the evaluation, a dominant factor is often existed. Through the study of the relationship between the internal structure of the original variable correlation matrix, several comprehensive indicators affecting academic early warning are found. In experiments, the comprehensive indicators are linear fitting of the original variables. The principal component analysis of the study is as follows:

Using statistic analysis tool, the principal component analysis was carried out on the indexes of "number of gates, total credits, obtaining credits, failing credits, passing rate, arithmetic average score, weighted average credit score, average credit score point, credit score point, failing score, number of library books borrowed and number of Library entrance". The analysis results are shown in Table II. It shows that the cumulative contribution rate from principal component 1 to principal component 4 has reached 86%, which can well reflect the characteristics of the whole sample.

In Table III, arithmetic mean, credit score, and average credit hour point have great influence on the first principal component (Comp1). Total credits and credits obtained have a greater impact on the second principal component (Comp2). Failing grade and rate of passing have a greater impact on the third principal component (Comp3). The library leading has a greater impact on the fourth principal component (Comp4). From above analysis, it is determined that the arithmetic mean, credits obtained, the average score point, and the library leading are the key factors that have a greater impact on the early warning of students' academic performance.

TABLE II. RESULTS OF PRINCIPAL COMPONENT ANALYSIS

Principal components/correlation — Number of obs = 794; Number of comp. = 11; Trace = 12; Rotation: (unrotated = principal); Rho = 1.0000

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.46996	2.90401	0.4558	0.4558
Comp2	2.56595	1.25041	0.2138	0.6697
Comp3	1.31554	.335884	0.1096	0.7793
Comp4	.979657	.123282	0.0816	0.8609
Comp5	.856375	.39021	0.0714	0.9323
Comp6	.466165	.167679	0.0388	0.9711
Comp7	.298486	.263639	0.0249	0.9960
Comp8	.0348468	.0289121	0.0029	0.9989
Comp9	.00593472	.000159245	0.0005	0.9994
Comp10	.00577547	.00447059	0.0005	0.9999
Comp11	.00130488	.00130488	0.0001	1.0000
Comp12	0	.	0.0000	1.0000

TABLE III. CORRELATION ANALYSIS OF PRINCIPAL COMPONENT

Variables	Comp1	Comp2	Comp3	Comp4
Number of courses	0.0265	0.4907	-0.0544	0.1869
Total credits	0.0769	0.5919	-0.0302	-0.0033
Credits obtained	0.1752	0.5356	-0.1803	0.0328
Failing grade	-0.3195	0.1331	0.4811	-0.1150
Rate of passing	0.3202	-0.0970	-0.4868	0.1256
Arithmetic mean	0.4005	-0.0876	0.2356	-0.0428
Credit score	0.4061	-0.0721	0.2269	-0.0604
Ave. credit hour points	0.3879	-0.0531	0.3071	-0.0691
Grade point	0.3752	0.1961	0.2581	-0.0736
Failing number	-0.3535	0.1447	0.0287	0.0585
Library lending	-0.1234	0.1048	0.4379	0.0378
Library returning	0.0355	-0.0937	0.1864	0.9563

C. Choose the optimal classification algorithm.

In this paper, the comprehensive data such as educational administration data and library data are pre-processed, and then the principal component analysis is used to analyze the factors affecting academic early warning. Four key factors are selected in the end. In each semester (term), only arithmetic average score, weighted average score and average score point are selected. However, the number of library borrowing is removed. After that, new research sample data are formed. Python language is used to model and analyze different algorithms in machine learning fields, such as decision tree, Bayesian, and artificial neural network. These algorithms are implemented and compared for accuracy of early warning classification.

Seventy percent of the data in the study sample is used as the training set, and thirty percent as the test set.

1) Decision tree

The decision tree is based on the probability of occurrence of various situations. It constructs a tree structure, and calculates the probability that the expected value of NPV is greater than or equal to zero, so as to evaluate the project risk. This is a decision analysis method to determine its feasibility, and is a graphical method for intuitively using probability analysis. Because this decision branch is shaped like a branch of a tree, it is called decision tree. Decision tree can be used as a prediction model, which represents a mapping relationship between object attributes and object values. Here, entropy is defined as the messy degree of the system. Algorithms for ID3, C4.5 and C5.0 spanning tree can be implemented to use entropy.

This measure is based on the concept of entropy in informatics theory. Generally speaking, decision tree is a very common supervised learning classification method. Regulatory learning is defined as a given set of samples, each of which has a set of attributes and a pre-determined category. A classifier can be obtained by machine learning, and it is able to give correct classification to new objects. The test is implemented by ID3 algorithm.

The Table IV is the accuracy rate of using the decision tree to test the 1-8 semester sample data. A total of 20 tests were conducted, and the last one was the average accuracy of the 20 tests. Among them, Term 1 (the first row) means that only the first semester of the study sample and the number of Library entries are tested. Term 2 indicates the first and second semesters of the sample and the number of Library entries. The rest are analogous.

TABLE IV. PREDICTION ACCURACY OF DECISION TREE

Term	Test 1	Test 2	Test 3	Test 4	Test 5	...	Average accuracy rate
1	0.83	0.82	0.8	0.84	0.85	...	0.837
2	0.87	0.83	0.85	0.83	0.9	...	0.8635
3	0.82	0.83	0.89	0.87	0.89	...	0.8615
4	0.85	0.89	0.88	0.89	0.85	...	0.86
5	0.84	0.83	0.86	0.88	0.87	...	0.855
6	0.84	0.86	0.87	0.83	0.87	...	0.858
7	0.84	0.8	0.85	0.86	0.89	...	0.865
8	0.84	0.86	0.79	0.86	0.89	...	0.872

2) Artificial neural network

The artificial neural networks (ANN) system were appeared after 1940s. It is composed of many neurons with adjustable connection weights. And it has the characteristics of large-scale parallel processing, distributed information storage, with good self-organizing and self-learning ability. Back propagation (BP) algorithm, also known as error back propagation algorithm, is a supervised learning algorithm in artificial neural network. BP neural network algorithm can approximate any function theoretically, and the basic structure is composed of nonlinear change elements, which have strong nonlinear mapping ability. The parameters of the network, such as the number of intermediate layers, the number of processing units in each layer and the learning coefficient of the network, can be set according to the specific conditions. It has a wide application prospect in many fields, such as optimization, signal processing, pattern recognition, intelligent control, fault diagnosis, and so on.

The following Table V uses artificial neural network to test the accuracy of the 1-8 semester data.

TABLE V. PREDICTION ACCURACY OF ANN

Term	Test 1	Test 2	Test 3	Test 4	Test 5	...	Average accuracy rate
1	0.83	0.84	0.73	0.72	0.8	...	0.7775
2	0.85	0.73	0.73	0.82	0.87	...	0.773
3	0.73	0.87	0.81	0.79	0.76	...	0.7615
4	0.76	0.77	0.72	0.88	0.71	...	0.751
5	0.78	0.71	0.8	0.78	0.77	...	0.7605
6	0.84	0.73	0.82	0.74	0.8	...	0.761
7	0.8	0.81	0.83	0.7	0.72	...	0.7695
8	0.74	0.77	0.73	0.76	0.73	...	0.7595

3) Bayesian classification algorithm

Bayesian classification algorithm is a statistical classification method, a kind of algorithm using probability and statistics knowledge. Data mining is based on Bayesian theorem. Naive Bayesian classification and Bayesian belief network are used for classification. Naive Bayesian classification assumes that the effect of an attribute value on a given class is independent of the value of other attributes. That is, there is no dependency between attributes, so it is called "naive". It is characterized by the expression of all forms of uncertainty in the form of probability. Both learning and reasoning are realized by probabilistic rules. The result of learning can be interpreted as the degree of trust in different possibilities. In lots cases, Native Bayesian (NB) classification algorithm are commonly compared with decision tree and neural network classification method. The algorithm can be applied to large databases, and the method is simple, accurate, and fast [8].

The following TABLE VI uses Native Bayesian algorithm to test the accuracy of the 1-8 semester data.

TABLE VI. PREDICTION ACCURACY OF NATIVE BAYESIAN

Term	Test 1	Test 2	Test 3	Test 4	Test 5	...	Average accuracy rate
1	0.85	0.84	0.85	0.85	0.83	...	0.854
2	0.88	0.9	0.88	0.84	0.89	...	0.878
3	0.86	0.87	0.88	0.89	0.89	...	0.8915
4	0.9	0.89	0.9	0.86	0.88	...	0.89
5	0.85	0.88	0.88	0.87	0.88	...	0.8805
6	0.88	0.9	0.9	0.88	0.87	...	0.882
7	0.85	0.89	0.88	0.87	0.86	...	0.883
8	0.89	0.87	0.89	0.89	0.87	...	0.8825

From data, it can be concluded that the average prediction accuracy of Bayesian algorithm is the highest among these three machine learning algorithms, suitable for AEWS project.

D. Determine prediction time for early academic warning

As for the specific time of early academic warning, there is little difference between the early warning results of sophomores (the third semester) and the later from the test data. In general, early warning results of the sophomores can be used to evaluate and judge the students' learning situation more steadily. The AEWS also provides tools to judge whether formal academic warning is necessary for a certain student. From TABLE III, the prediction accuracy of the first semester is 85.4% by using Naive Bayesian algorithm. If we can use the data at the end of the first semester to find out the students with potential learning difficulties and intervene in time, the effect is much better than that of intervening again after sophomore or junior year.

IV. CONCLUSION

In this paper, we collect comprehensive data of students from various departments, such as the academic affairs office, library, and other departments. The principal components analysis was used to find out the key predictors after pretreatment. Three classic machine learning classification algorithms for data mining are implemented to train and test sample data. Finally, Bayesian algorithm is selected as the best academic warning model. The results show that the Naive Bayesian algorithm, based on the first three semesters of students' academic affairs and library related data, can provide more accurate results. The accuracy rate remains above 86%. At the same time, the accuracy rate of the algorithm at the end of the first semester is as high as 85.4%. At this point, helping students with predicted academic difficulties can effectively reduce the possibility that students will be formally warned later. In recent years, about 160,000 college students in China drop out of school every year, accounting for 0.75% of the students. It is necessary for domestic universities to implement early warning system, which plays a supervisory role for students, unify parents, teachers and students. Through the establishment of academic monitoring, early warning system is able to help more students successfully complete their studies [9,10]. This system can be applied to other colleges and universities with minor modification. Furthermore, the system model can also be extended to other personalized management fields, such as student group analysis, learning goal prediction and so on, so as to make a wider application to the comprehensive analysis and mining of large data of higher education. And it also effectively serve the decision-making and management of colleges and universities.

ACKNOWLEDGMENT

This research was financially supported by the scientific research funds of Hangzhou Normal University, and its algorithms is implemented in the school of information science and engineering [11,12,13].

REFERENCES

[1] S. P. Xiao, “Exploration of College Students' Academic Early Warning Mechanism,” Contemporary Education Forum of Management Research, vol. 8, pp.14-16, 2011.

[2] B. L. Yang, Z. H. Guo, “Construction of College Students' Academic Early Warning Mechanism for Promoting Professional Development,” Software Guide: Educational Technology,vol. 9, pp.35-37, 2012.

[3] H. Y. Zhang, “Early Warning Mechanism for Learning in Colleges and Universities,” Science and Technology Information, vol., pp.801-806, 2010.

[4] B. Fan, “Practical Exploration of Academic Early Warning Mechanism in Higher Vocational Colleges,” Journal of Changzhou Institute of Information Technology, vol. 4, pp.73-75, 2011.

[5] Z. H. Wang, Luo Baoqing, “Analysis of the Causes of College Students with Learning Difficulties and Construction of Early Warning Mechanism,” People's Forum,vol. 8, pp.172-173, 2011.

[6] R. Balfanz, J. Fox, “Early Warning Systems-Foundational Research and Lessons from the Field,” National Governors Association, Washington D. C., 2011.

[7] Y. F. Zhang, R. Hu, “Multivariate Comprehensive Evaluation Method of Principal Component Analysis Model,” Journal of Southwest University for Nationalities, vol. 39 (3), pp. 362-365, 2013.

[8] C. Zhang, M. L. Guo, “Improvement and Implementation of Naive Bayesian Classification Algorithm in Large Data Environment,” Journal of Beijing Jiaotong University, vol. 39 (2), pp.35-41, 2015.

[9] D. P. Yang, “Report on China's Education Development 2012,” Social Science Literature Press, Beijing, 2012.

[10] Z. Y. Zhao, J. Sun, Z. Y. Jiang, “Promotion of Academic Early Warning System in Applied Undergraduate Universities from the Perspective of Internet ,” Henan Science and Technology, vol. 11, pp. 265-266, 2015.

[11] B. Wang, W.L. Zhou, “Comprehensive Integrated Platform for Garbage Classification in Reduction Innocuity, and Resource,” Int. Conf. on Computer, Communication and Network Technology, pp. 105-111, 2018.

[12] B. Wang, W.L. Zhou, Zhihua Li, “Research on Distributed Intelligent Mattress on the Internet of Things,” Int. Conf. on Computer Science and Information Engineering, pp. 84-88, 2018.

[13] B. Wang, W. L. Zhou, S. H. Shen, “Garbage Classification and Environmental Monitoring based on Internet of Things,” Int. Conf. on Information Technology and Mechatronics Engineering Conference , pp. 23-27, 2018.