Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 3, March 2019, pp. 822-833
DOI: 10.23940/ijpe.19.03.p11.822833

Student Performance Early Warning based on Data Mining

Chunqiao Mi[a,b,*]

[a]*School of Computer Science and Engineering, Huaihua University, Huaihua, 418000, China*
[b]*Key Laboratory of Intelligent Control Technology for Wuling-Mountain Ecological Agriculture in Hunan Province, Huaihua, 418000, China*

Abstract

Student performance in higher education is related to many complicated factors and always has uncertainty, so early warning of it is a very difficult issue. In this study, a systematic review was first carried out on student performance prediction and early warning using data mining techniques, including basic data sources, evaluating factors, predicting methods, application tools, and practices. Then, insufficiencies of the related studies were discussed, including incomprehensive source data, inadaptable and unspecialized calculation methods, and lack of integrated methodology systems in practice. Finally, a solution design was proposed, consisting of learning situation big data, a systematic early warning model, and an integrated information support system. Preliminary experiment results showed that it could identify at-risk students in a timely manner and improve the overall efficiency and effectiveness of early warning education management in practice, so it is of both academic and practical significance in promoting the deep integration of information technology and early warning education.

Keywords: student performance early warning; data mining techniques; learning situation big data; artificial neural network; integrated methodology system

(Submitted on October 23, 2018; Revised on November 25, 2018; Accepted on December 21, 2018)

1. Introduction

At present, as the higher education enrolment expansion scale becomes larger and larger, students' academic quality has begun to drop sharply; therefore, early warning of student performance is a basic need in higher education management. However, predicting student performance as early as possible is very challenging due to the following: firstly, a student's academic failure is related to many factors and always has uncertainty, also known as the "one thousand factors problem" [1]; secondly, there is limited systematic investigation on the methods that are suitable and effective for evaluation and early warning of student performance; lastly, the lack of integrated information systems for monitoring, analyzing, and predicting student study progress and performance is currently not being addressed [2]. Aiming at these problems, in this study, a systematic review on prediction and early warning of student performance based on data mining techniques was firstly carried out, and then a solution method that can identify at-risk student in a timely manner was proposed. It can provide a good reference for improving early warning education management.

2. Current Research Status

2.1. The Data Used

With the rapid development and further application of modern information tools and systems in every stage of higher education, there is now a large amount of educational process data available to us, and the two most used data resources in assessing student performance are online and offline learning data, as shown in Table 1. Online learning data is usually obtained from different online study support systems like e-learning and learning management systems, adaptive educational hypermedia systems, and intelligent tutoring systems [3]. At the moment, it is this type of data that plays the core role in quantitative analysis of student study performance. Offline learning data, such as some evaluated result data, is

* Corresponding author.
E-mail address: michunqiao@163.com

usually drawn from traditional exams, tests, assignments, and seminars taken in or out classes [4-5], which are usually used to assess students' learning performance especially in a specific subject's evaluation. Both of these educational data resources are very helpful for teachers to record students' learning behaviours and help them improve their academic performances; however, at present, these data resources are mainly about students themselves rather than teachers or the interaction between students and teachers, while the latter is also very important for student performance formation and should be included in student performance early warning.

Table 1. The main data resources used in student performance early warning

Data category	Data name	Data source example
Online learning data	Learning process data	E-learning and learning management systems, etc.
	Learning result data	Educational administration management systems, etc.
Offline learning data	In-class data	In-class exams, tests, assignments, etc.
	Out-class data	Out-class seminars, homework, experiments, etc.

2.2. The Evaluating Factors

According to our systematical review on related literature, four mostly used factors used in evaluating student performance were sorted out and are shown in Table 2. The first one is the inter-curricular index, which is the most frequently used, including attributes such as assignments, quizzes, lab work, class tests, and attendance [6-8]. These attributes are usually considered as indications of student personal realized academic potential, which has a tangible value for future career success [9]. They are always used as the most influential factors in determining the survival of a student in his or her studies [10]. The second one is the extra-curricular index, including attributes like final grades obtained in particular subjects and extracurricular activities [6-7,9], which can directly reflect student learning outcomes and learning process effectiveness. The third is the student demographic index, including information about student gender, age, disability, family background, high school background, and so on [7,9-10]. Different attributes can lead to different learning behaviors and styles. For example, compared to male students, it was found that most female students are usually more self-directed, dutiful, and focused on their studies, so as to yield distinct performance outcomes [11]. The fourth is the student social index, including student social relationship networks, participation in online discussion forums, textual communication data generated in the MOOCs, and family team support [12-13], which can imply a student's learning effort and status so as to indirectly affect student performance. However, at present, the attributes of the student psychological index, such as students' study motivation and interest, are rarely used in evaluating student performance. This may be because they are usually qualitative data and very difficult to be collected, but they are also very important in reflecting the inner state of students' learning activities.

Table 2. The factors used in evaluating student performance

Factor category	Index name	Attribute value example
Curricular related factor	Inter-curricular index	Cumulative grade point average, assignment, quizzes, lab work, class test, attendance, etc.
	Extra-curricular index	Final grades obtained in particular subjects, extracurricular activities, etc.
Background related factor	Student demographic index	Gender, age, disability, family background, and high school background, etc.
	Student social index	Student social relationship network, participation in online discussion forum, etc.

2.3. The Related Methods

In predicting students' performance and behavior towards their studies, statistics as well as machine learning provide various methods to extract valuable information from a variety of available educational data. Different methods have different characteristics, which can be sorted out as shown in Table 3, in which the top five are the most used methods with relatively higher accuracy [2].

Some descriptively statistical methods have been applied to investigate student performance related issues. For example, logistic regression was used by Baker et al. and Jay et al. to find early indicators of student success or failure [14-15]. A method combining factor analysis and logistic regression was used by Campbell, Bravo et al., and Leah et al. to detect underperforming students [16-18]. The method of discriminant analysis was used by Minaei et al. and Morris et al. to predict course success and student scores [19-20]. Principal component analysis and multiple linear regressions were used by Yang et al. to predict student academic performance [21]. These related studies have made progress in identifying at-risk students and early warning of student performance. However, most of these attempts are still focused on the analysis of students' knowledge point mastery for early warning, so the warning indicators are not comprehensive enough. Additionally, the methods used usually involve the analysis of static result data rather than dynamic process data. Therefore, the obtained final results are mostly descriptive rather than predictive, and the accuracy is usually not very high.

On the other hand, there are also some predictive methods based on machine learning algorithm showing new data-driven features. For example, interaction data was used by Douglas et al. and Geraldine et al. to predict student failing risk [22-23]. Predictive models were reported by Barber and Sharkey to identify academically at-risk students [24]. Association rules and decision trees were used by Chen et al., Ma et al., and Bravo et al. to identify potential low-performing students [17,25-26]. The artificial neural network was applied by Mi et al. and Yang et al. to predict student grade and learning failure risk [27-28]. The Bayes approach was used by Kris et al., Madhavi et al., Mollica et al., and Hamoud et al. to determine student study performance [29-32]. A classification system was used by Martin et al. and Kevin et al. to identify poor performers during current dynamic course studying [33-34]. Support vector machine was used by Sandeep et al. to early detect academically difficult undergraduate students during the course studying progress [35]. These new attempts have promoted the deep integration of data mining techniques and early warning education, but the number of initiatives that have been able to transition from algorithm concept to practice application is still scarce; moreover, there is also a lack of integrated methodology system solutions.

Table 3. The related methods used in assessing student performance

Name	Characteristic	Category
Decision tree	is simple and comprehensible to uncover small or large data structure and predict the value	machine learning, supervised, predictive
Neural network	can detect all possible interactions between dependent and independent variables, especially model complex nonlinear relationship	machine learning, supervised, predictive
Naive bayes	is easy to use all of attributes contained in the data	statistic method, supervised, predictive
K-Nearest neighbor	can estimate the detailed pattern for learner's progression	machine learning, unsupervised, predictive
Support vector machine	is suited well in small datasets, has a good generalization ability	machine learning, supervised, predictive
Association rule	can discover hidden associations or relationships between data items	machine learning, unsupervised, descriptive
Multiple linear regression	is simple and convenient to analyze the multi factor problem	statistic method, supervised, predictive
Logistic regression	prediction result is a probability between 0 and 1, easy to explain and use.	statistic method, supervised, predictive
Discriminant analysis	discriminant function is easy to understand and use	statistic method, supervised, predictive
Factor analysis	can use few factors to describe the relationship between many indicators	statistic method, supervised, descriptive
Principal component analysis	can reduce dimensionality	statistic method, supervised, descriptive

2.4. The Tools and Practices

Due to the increasingly large amount of education-related data, in the current big data era it is impossible to extract information from big data sets by hand. In order to deal with this issue, various software tools, including open source ones and commercial ones, are coming into existence and are available to us. For example, the Java-based open source tool WEKA is one of the most used data mining tools, with which you can easily implement your own algorithms and obtain your results very quickly [36]. The RapidMiner is another very powerful data mining software tool. Using tasks such as business analysis, machine learning, and even text mining can easily be done in only one integrated platform, and it is also very convenient to finish the tasks of data preprocessing like classifying the whole data into testing, training samples, and performing ten-fold validation [37-38]. Some commercial tools such as SAS and SPSS Modeler also play important roles in data processing and model implementing. Although these general tools can greatly facilitate educational data mining, in the context of student performance prediction and early warning, most of these tools usually can only be used in scientific research rather than in education management practice.

Fortunately, there have been also some initial practical attempts on student performance assessment and early warning. For example, based on student study process data, the Course Signal and Purdue Signal projects at Purdue University can give students their curriculum performance information in a timely manner and identify at-risk students with the Student Success Algorithm [39-40], which is very useful for students to adjust their study strategies in time and improve curriculum success rates. Schools such as the University of Alabama, Arizona University, University of Phoenix, and University of Maryland have also carried out some related practices using multi-source educational data to predict student study crisis, which is very beneficial for students to enhance their academic success and for schools to keep high student retention [35]. In addition, the Student Success System developed by the Desire2Learn agency [41], the Study Dashboard Platform developed by Khan Academy [42], the Starfish Early Alert System in the Starfish Enterprise Success Platform, and the early warning application system developed by the Education Data Research Institute in University of Electronic Science and Technology in China [43] all have some practical early warning functions with good effects in reducing student academic risk and improving overall education quality. However, these initial practices usually just exist only as a part of other information systems rather than as a stand-alone integrated system, so it is difficult to improve the whole efficiency and effectiveness of early warning education practice.

3. Problem Analysis

3.1. The Data Used is Incomprehensive

Firstly, open data sets are not easily available. At present, although there are more and more data generated every day, public and open source data sets are still very rare for student performance assessment. Due to the consideration of privacy protection, economic cost, and academic significance, the data sets used in most literature are not available for readers. This will reduce the credibility and influence of the research results. A lack of open data sets will also hinder the development of student performance early warning research based on educational data mining.

Secondly, the evaluating factors are not comprehensive. A student's learning outcome is the result of combined effects of many external and internal factors. However, most of the current research only focuses on students' external behavior data, but little is known about their internal psychological behaviors such as student motivation, interest, attitude, engagement, family support, attention focusing, and so on, which is very important in predicting student learning status. For example, personal internal eye-tracking data is useful for the lecturer to understand students' focus based on their personal eye-tracking interest, and it can also be used to predict student performance and abnormal behavior combined with other external behavior data.

Lastly, the amount of data is still very small. The number of students involved in current studies is generally from dozens to hundreds, and the size of data sets is from several KB to dozens of MB. These cases cannot yet be regarded as educational big data research. Additionally, most of the previous research done in this field usually focuses on new open educational environments, like distance or online education, from which the data can easily be collected. Traditional closed educational environments, like the traditional classroom, have drawn less attention in this research [44]. However, they are still in an important position in today's education system. For example, the personal face-to-face interactions between teachers and students is the most helpful information for understanding students' problems. Therefore, there is still a need to pay attention to the data from traditional educational environments.

3.2. The Method Used is Inadaptable

Firstly, less attention is paid to data pre-processing on the problems of high dimension and imbalance. Most research usually pays more attention to the educational data mining algorithm itself, but it should be noted that data pre-processing is also a very important work, because the quality of input variables can directly affect the reliability of obtained results. In predicting student performance, there are many factors that can affect students' final outcomes. For example, in one research there is a total of 77 factors associated with student performance in school [1]. This is usually known as the high dimensionality problem, referring to the fact that the number of related attributes is very large especially in today's big data era. The data in the field of student performance analysis is usually imbalanced, because most students have a performance of success while only a minority of them fail. In this imbalanced situation, data mining algorithms would usually pay more attention to the majority class but overlook the minority one, which could lead to a poor result. Therefore, before applying educational data mining techniques, more attention should be paid to data pre-processing including data cleaning, filtering, transforming, partitioning, dimensionality reducing, and data rebalancing, so that the work of student performance assessment and early warning can be carried out more properly and accurately.

Secondly, most methods used are about static analysis of student learning results rather than dynamic predicting or warning of learning procedural risks, and the warning time is usually too late to provide appropriate pedagogic interventions. Most of the current research mainly focuses on assessment or prediction of static performance results rather than dynamic risk analysis along with student learning progress, and the risk reasons constitution analysis is also insufficient, so they can hardly provide substantial guidance for students' academic improvement. In the very early stage of student study, the information that can be collected and used is very little and always has uncertainty, so it is difficult to use conventional methods to perform early warning accurately. Therefore, the traditional methods usually must be implemented after the final examination or even later in the next semester in order to get maximum information and achieve better accuracy. In practice, it is too late for educators to detect potential at-risk students in advance or intervene to prevent them from getting worse in their studies.

Lastly, the methods used are more concerned about students than teachers, and there is a lack of adaptable methods that can be suitable for both the new online education environment and the traditional classroom environment. Student's learning performance is always obtained under the comprehensive effects of the student's own efforts and the teacher's rational guidance and interference, without any of which good results can hardly be reached. However, it is much more difficult to collect data on the teacher than on the student, so previous studies were mostly focused on analyzing and mining of student

data while less attention was paid to the effect of teachers. In reality, the role of teachers, including parents, in teaching and learning is also very important and should not be ignored; otherwise, it would be difficult to find the root risk causes of student performance and appropriate coping strategies for early warning. In addition, the open online learning environment is currently mainstream in learning performance prediction and early warning; they are closely integrated with modern information technology, from which it is easy to obtain data related to student learning. Therefore, many kinds of data mining and analysis methods are aimed at this new environment rather than the traditional classroom environment. However, the traditional environment is still an indispensable part of our education system, so it is very necessary to develop new methods that can be applied to both the new environment and traditional environment and can integrate data from both environments.

3.3. There is Lack of Integrated Methodology System in Practice

Firstly, there is currently no specialized data mining technique for education fields. With the rapid development of modern information technology, the research methods on student performance assessment and early warning have shown some new trends like data-driving and procedurization. However, these attempts are still only applications of the mature computer science or mathematic methods in the fields of education, and the number of initiatives that have been able to transit from concept to practice with modern information technology is still very scarce. For example, traditional statistical methods from mathematics, such as regression and discriminant analysis, were used most frequently in identifying factors and their contributions to student performance [45]; however, it would be hard or impossible to analyze due to the enormous volume of today's educational big data, within which the useful educational patterns exist [46]. Therefore, it is very urgent to develop new specialized data mining techniques that can deeply integrate information technology and education context.

Secondly, the practical tools used in analysis and prediction of student performance are very outdated. In order to detect students at high-risk of failure as early as possible, an early warning tool is a good solution, and efforts to apply it in education have increased in recent years and there are some practical examples [47]. Using specific critical thresholds and three indicators, an early warning tool based on a Microsoft Excel file was developed, and some guidelines for following the students' education were defined by the Mexico Sub-Secretary of Middle Education [47]. A guide and an early warning tool based on a template from Microsoft Excel [48] were also defined by the U.S. National High School Center. These practical tools can implement some tasks like identifying the factors with strong correlation to student dropout and prediction of students dropping out but given that they use simple Excel files by hand, they are very outdated and not appropriate when the amount of data becomes larger. Besides, the methods used in most current practices mainly focus on statistically static analysis of students' performance results data rather than dynamic analysis of students' learning process data. Therefore, the results obtained are mostly descriptive rather than predictive.

Lastly, there is still a lack of integrated information support systems for student performance early warning. In the domain of education, the final goal of performance prediction and early warning is to help students improve their learning outcomes, so an integrated early warning system should consist of a set of procedures and instruments for early detection of at-risk students, analysis of students' abnormal behaviors and statues, estimation of a students' knowledge, skill, and scores, and implementation of appropriate pedagogic interventions to help students overcome their weakness. However, most of the current research focuses on only one of these aspects rather than a comprehensive methodology system, and there is still very limited research on the design and construction of an integrated practical system. This is an important issue that must be addressed to improve early warning education practice.

4. Solution Design

4.1. The Overall Framework

Therefore, in order to deal with the existing problems, in this study a systematic early warning methodology solution was provided based on data mining techniques and learning analytics, as shown in Figure 1. It refers to the whole process of identifying students' academic crisis and predicting the specific risk level based on data mining related methods. The overall framework can be divided into the following three parts.

The first part was preprocessing calculation. In this step, all raw data were preprocessed and integrated as learning situation big data. For each student, the academic basis index was calculated as the weighted average of the source data of attendance, assignment, etc. The academic quality index and academic environment index were also calculated in the same way based on relevant source data shown in the top of Figure 1. The second part was student performance prediction. In this step, an artificial neural network prediction model was employed, which was composed of input and hidden and output layers. The above calculated academic basis index, academic quality index, and academic environment index were fed into

the input layer, and the whole prediction process included output value inferring, loss function evaluating, network training, and result validating. The third part was risk classification and visualization. In this step, the predicted result from the second part was divided into two categories (normal and abnormal) and four levels (no risk, low risk, medium risk, and high risk) based on both quantitative and qualitative methods including quantile analysis and expert experience. Multi-color-view visualization methods were also employed to represent the risk results, including risk evolution diagrams, comparative analysis diagrams, electronic dashboards, and traffic light warnings, which provided useful and easily understood risk information to students, teachers, college administrators, and parents through the integrated early warning information support system.

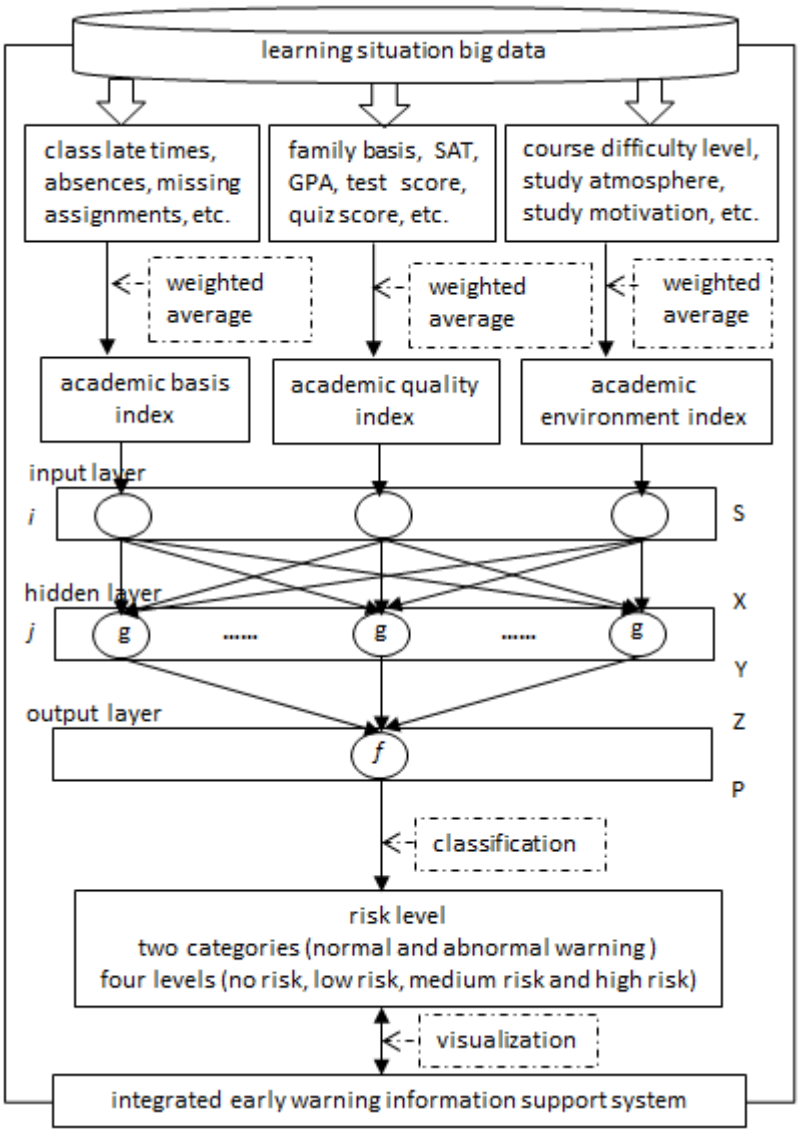

Figure 1. The overall framework of systematic early warning model

4.2. The Learning Situation Big Data

In the learning situation big data integration solution shown in Figure 2, at first, all relevant data about the student learning process, learning participation, learning interaction, learning performance, learning experience, learning environment, teaching process, education management process, and student demographic characteristics were comprehensively collected from sources such as the education administration management system, teaching assistant system, student online study system, mobile learning support terminal, and questionnaire surveys. Then, all kinds of raw data were desensitized and preprocessed by data cleaning, outlier detecting, missing value repairing, metadata tagging, etc., and all the original values were also normalized to values between 0 and 1. Finally, all data were integrated together according to the unique student ID, so as to form student learning situation big data. This student-centered learning situation big data is the root foundation

of automatically identifying and alerting student study failure risk in a timely manner.

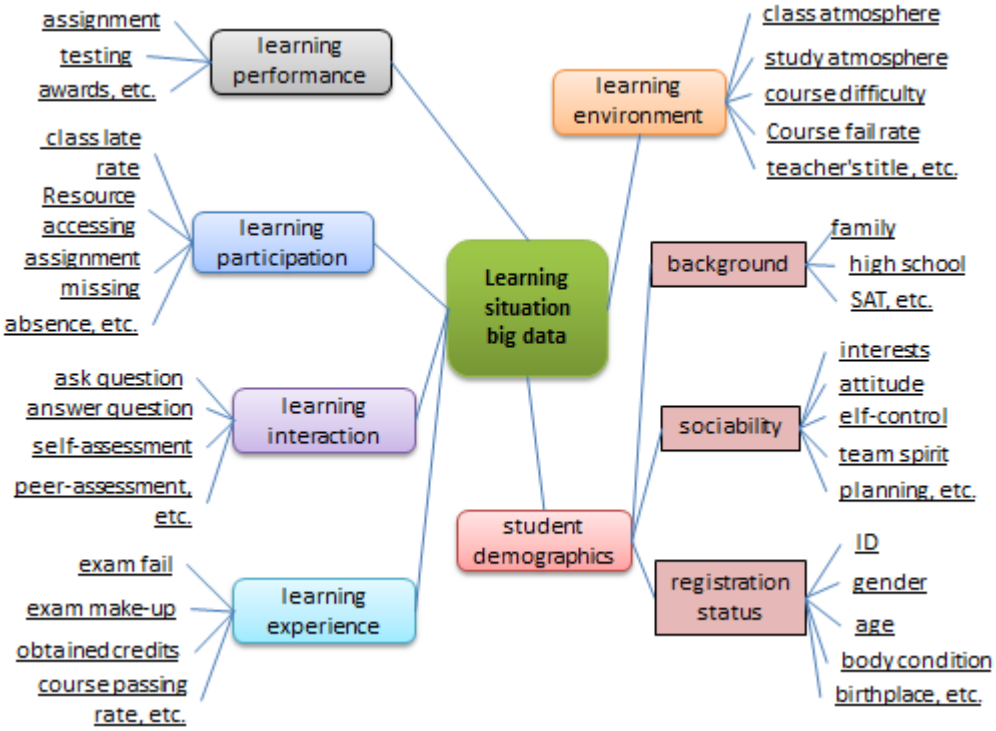

Figure 2. The integration of student learning situation big data

4.3. The Calculation Method

In our artificial neural network prediction model based on the backward propagation algorithm, there are three layers: the input layer, hidden layer, and output layer. In the input layer, there are three neuron nodes accepting the values of academic basis index, academic quality index, and academic environment index respectively, and we call the node in the input layer inputNode*i* ($1 \le i \le N$, $N = 3$). In the hidden layer, we assume the number of neuron nodes is M, and we call the node in the hidden layer hiddenNode*j* ($1 \le j \le M$). In the output layer, there is only one neuron node, and we call it outputNode. The calculation process includes two main stages, namely the information forward propagation stage and the error backward propagation stage.

In the first stage, the prediction value can be calculated by Equations (1) to (5). In Equation (1), S^{α} is the α^{th} input sample with N attributes corresponding to the neuron nodes in input layer, $1 \leqslant \alpha \leqslant Q$, and Q is the sample number. In Equation (2), X_{xj}^{α} is the input value of hiddenNode*j*, $S_{\alpha i}$ is the i^{th} attribute value of S^{α}, w_{ij} is the weight connecting inputNode*i* and hiddenNode*j*, and b_j is the bias of hiddenNode*j*. In Equation (3), $Y_{\alpha j}$ is the output value of hiddenNode*j* and g is the sigmoid activation function of the hidden layer. In Equation (4), Z^{α} is the input value of outputNode, w_j is the weight connecting hiddenNode*j* and outputNode, and c is the bias of outputNode. In Equation (5), P^{α} is the output value of outputNode corresponding to our prediction value and f is the pure linear activation function of the output layer.

$$S^{\alpha} = (S_1^{\alpha}, S_2^{\alpha}, \cdots, S_N^{\alpha}) \tag{1}$$

$$X_j^{\alpha} = \sum_{i=1}^{N} w_{ij} \times S_i^{\alpha} + b_j \tag{2}$$

$$Y_j^{\alpha} = g(X_j^{\alpha}) = g(\sum_{i=1}^{N} w_{ij} \times S_i^{\alpha} + b_j) \tag{3}$$

$$Z^{\alpha} = \sum_{j=1}^{M} w_j \times Y_j^{\alpha} + c = \sum_{j=1}^{M} w_j \times g(\sum_{i=1}^{N} w_{ij} \times S_i^{\alpha} + b_j) + c \tag{4}$$

$$P^{\alpha} = f(Z^{\alpha}) = f(\sum_{j=1}^{M} w_j \times g(\sum_{i=1}^{N} w_{ij} \times S_i^{\alpha} + b_j) + c) \tag{5}$$

In the second stage, in order to get the prediction value much closer to the actual observed value, the weights and biases should be adjusted. Therefore, during the network training process with sample-by-sample training strategy, the adjustments of the weights could be calculated by Equations (6) to (10) using the gradient descent method. In Equation (6), E^{α} is the error function also known as the loss function and O^{α} is the actual observed result value corresponding to the αth input sample. Equations (7) and (8) show the weight adjustment between hiddenNode*j* and outputNode, and η is learning rate, which is a positive constant in the range (0, 1). Equations (9) and (10) show the weight adjustment between inputNode*i* and hiddenNode*j*. The adjustments of the biases can be calculated in the same manner as those of the weights.

$$E^{\alpha} = \frac{1}{2}(O^{\alpha} - P^{\alpha})^2 \tag{6}$$

$$-\frac{\partial E^{\alpha}}{\partial w_j} = -\frac{\partial E^{\alpha}}{\partial P^{\alpha}} \times \frac{\partial P^{\alpha}}{\partial Z^{\alpha}} \times \frac{\partial Z^{\alpha}}{\partial w_j} = (O^{\alpha} - P^{\alpha}) \times f'(Z^{\alpha}) \times Y_j^{\alpha} \tag{7}$$

$$\Delta w_j^{\alpha} = \eta \times (-\frac{\partial E^{\alpha}}{\partial w_j}) = \eta \times (O^{\alpha} - P^{\alpha}) \times f'(Z^{\alpha}) \times Y_j^{\alpha} \tag{8}$$

$$\begin{aligned} -\frac{\partial E_{\alpha}}{\partial w_{ij}} &= -\frac{\partial E^{\alpha}}{\partial P^{\alpha}} \times \frac{\partial P^{\alpha}}{\partial Z^{\alpha}} \times \frac{\partial Z^{\alpha}}{\partial Y_j^{\alpha}} \times \frac{\partial Y_j^{\alpha}}{\partial X_j^{\alpha}} \times \frac{\partial X_j^{\alpha}}{\partial w_{ij}} \\ &= (O^{\alpha} - P^{\alpha}) \times f'(Z^{\alpha}) \times w_j \times g'(X_j^{\alpha}) \times S_i^{\alpha} \end{aligned} \tag{9}$$

$$\Delta w_{ij}^{\alpha} = \eta \times (-\frac{\partial E^{\alpha}}{\partial w_{ij}}) = \eta \times (O^{\alpha} - P^{\alpha}) \times f'(Z^{\alpha}) \times w_j \times g'(X_j^{\alpha}) \times S_i^{\alpha} \tag{10}$$

In order to optimize network parameters like node number in the hidden layer, learning rate, and so on, at first they were initialized as random values, and then the relative root mean square error measurement and the method of trial and error were employed to test and optimize them. After the optimization and training process, the parameters and network topology can be determined and the prediction value can be obtained when a new sample is fed into the network. Then, the final risk level is determined by Equation (11), where *P* is the prediction value and *NR*, *LR*, *MR*, and *HR* mean no risk, low risk, medium risk, and high risk respectively.

$$RiskLevel = \begin{cases} NR, \text{ if } 1 \geq P \geq 0.6 \\ LR, \text{ if } 0.6 > P \geq 0.4 \\ MR, \text{ if } 0.4 > P \geq 0.2 \\ HR, \text{ if } 0.2 > P \geq 0 \end{cases} \tag{11}$$

4.4. The Prototype System

Based on the above methods, a prototype integrated early warning information support system has been developed by us, in which there are five architecture layers including the data layer, algorithm layer, application layer, environment layer, and user layer. In the data layer, the basic learning situation related data were stored in the database as well as file management systems. In the algorithm layer, the related models were implemented using C# language in the background of the system. In the application layer, the risk information query and display were realized with HTML5, CSS3, and JavaScript technologies. In the environment layer, the end-user environment can be adapted to both the new online environment and traditional classroom environment with adaptive use of PC, tablet, and mobile phone as terminals. In the user layer, timely and accurate learning performance-related information can be provided for students, teachers, college administrators, and parents. With this integrated information support system, early warning of student performance can be carried out step by step automatically during the continuous progress of learning activities. Once a student is identified with academic risk at a certain stage, the system can immediately send warning signals to the relevant students, so as to help students adjust

learning methods and allow teachers to provide pedagogic interventions in time. This can assist students in mastering their learning crisis at any time and adjusting learning methods and strategies in a timely manner, so as to improve study success rates.

4.5. The Preliminary Result and Discussion

Firstly, a data set including 400 students was used to train and test our prediction model. This data set was randomly divided into two sub data sets, training data set and testing data set, including 300 and 100 students respectively. The accuracy measurements of these two data sets are shown in Table 4, where it can be seen that the error of the training data set was a bit smaller than that of the testing data set. However, in general, the errors were both relatively small in the education domain, indicating that the obtained prediction model was promising.

Table 4. The training and testing accuracy measurements of our model

Data set	Sample number	Relative root mean square error
Training	300	0.1612
Testing	100	0.1553

Then, with the prototype system, an initial experiment was carried out on 225 new students. Preliminary results showed that 82 of them were identified as at-risk students who had some academic problems in different aspects and different stages during their studying process. Among them, there were 13, 28, and 41 students falling into high, medium, and low risk levels, as shown in Figure 3. It can also be see that some of them obtained improvement in their studies after warning signals and learning advice were provided in a timely manner. However, it should also be noticed that the improvement percentages of different risk levels are not the same, as they are 23%, 67%, and 80% respectively. Therefore, early warning information has the least stimulating effect on high-risk students and the greatest effect on low-risk students. This also tells us that in education practice, we should direct more remedial measures towards high-risk students in order to rescue their academic failure.

Under the help of the integrated information support system, we can easily pay attention to all observed students rather than only the students with problems; trace students' potential academic crises from attendance, homework, in-class tests, and other issues rather than just from students' previous course scores; provide early warnings in real time during the learning process rather than after the end of the course; and send early warning messages to stakeholders automatically through the system rather than make telephone calls by hand. Additionally, the number of staff involved in early warning education management can also be reduced from more than ten to only one. Therefore, the solution method provided is both effective and efficient in practice.

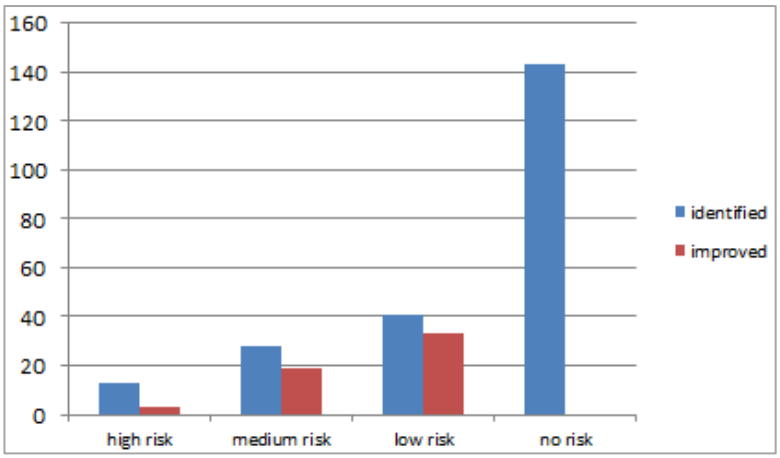

Figure 3. The predicted risk results and the students with improvement

5. Conclusion and Future Work

Student performance early prediction and warning is a serious problem in education, and it is also not an easy issue to resolve. Through a systematical review on related literature in this study, it is determined that in the early warning of student performance, there are two main data sources including online learning data and offline learning data, four usually used indices including the inter-curricular index, extra-curricular index, student demographic index, and student social interaction index, five mostly used prediction methods including decision tree, neural network, naive Bayes, k-nearest neighbor, and support vector machine, and two main types of practical tools including open source tool and commercial tool. However, in

practice there are still some insufficiencies, such as not comprehensive data, a not adaptable method, and a lack of integrated information support methodology systems. Aiming at these problems, we proposed a solution consisting of learning situation big data integration, a systematic early warning model, and an integrated information support system based on data mining techniques. Preliminary experiment results showed that it can identify at-risk students in a timely manner before the final examination with comprehensive, procedural, and dynamic characteristics. It has academic significance and practical significance, as it can improve the efficiency and effectiveness of early warning education management.

However, in terms of student performance early warning, we are not satisfied with just the identification of at-risk students and performance prediction. The individualized pedagogical interventions or learning guidance prescriptions should also be proposed based on accurate evaluation of the at-risk students' weaknesses, so as to help the students defuse their study crisis and improve their overall study success. Therefore, as a next step, the academic improvement intervention model and methods will be further studied based on data mining techniques.

Acknowledgements

We are very grateful that our study is supported by the Hunan Provincial Philosophy and Social Sciences Foundation (No. 17YBQ087), the Program of Hunan Provincial Social Science Achievements Evaluation Committee (No. XSP18YBC182), the Hunan Provincial Educational Science 13th Five-Year Planning Program (No. XJK016QXX003), the Hunan Provincial Natural Science Foundation (No. 2017JJ3252), and the teaching reform project "Research on the individualized teaching reform of software engineering major under the background of new engineering".

References

1. C. Márquez-Vera, A. Cano, C. Romero, and V. Sebastián, "Predicting Student Failure at School using Genetic Programming and Different Data Mining Approaches with High Dimensional and Imbalanced Data," *Applied Intelligence*, Vol. 38, No. 3, pp. 315-330, 2013
2. A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques," *Procedia Computer Science*, No. 72, pp. 414-422, 2015
3. L. Razzaq, J. Patvarczki, S. F. Almeida, M. Vartak, M. Feng, N. T. Heffernan, et al., "The Assistment Builder: Supporting the Life Cycle of Tutoring System Content Creation," *IEEE Transactions on Learning Technologies*, Vol. 2, No. 2, pp. 157-166, 2009
4. S. T. Hijazi and S. M. M. R. Naqvi, "Factors Affecting Students' Performance a Case of Private Colleges," *Bangladesh E-Journal of Sociology*, Vol. 3, No. 1, pp. 1-10, 2006
5. S. K. Yadav, B. Bharadwaj, and S. Pal, "Data Mining Applications: A Comparative Study for Predicting Students' Performance," *International Journal of Innovative Technology & Creative Engineering*, Vol. 1, No. 12, pp. 13-19, 2011
6. D. M. D. Angeline, "Association Rule Generation for Student Performance Analysis using Apriori Algorithm," *The SIJ Transactions on Computer Science Engineering & its Applications*, Vol. 1, No. 1, pp. 12-16, 2013
7. S. Natek and M. Zwilling, "Student Data Mining Solution-Knowledge Management System Related to Higher Education Institutions," *Expert Systems with Applications*, Vol. 41, No. 14, pp. 6400-6407, 2014
8. S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving Accuracy of Students Final Grade Prediction Model using Optimal Equal Width Binning and Synthetic Minority Over-Sampling Technique," *Decision Analytics*, Vol. 2, No. 1, pp. 1-25, 2015
9. U. B. Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An Overview of using Academic Analytics to Predict and Improve Students' Achievement: A Proposed Proactive Intelligent Intervention," in *Proceedings of the IEEE 5th International Conference on Engineering Education*, pp. 126-130, Selangor, Malaysia, 2013
10. T. M. Christian and M. Ayub, "Exploration of Classification using NB Tree for Predicting Students' Performance," in *Proceedings of the International Conference on Data and Software Engineering*, pp. 1-6, Bandung, Indonesia, 2014
11. A. Simsek and J. Balaban, "Learning Strategies of Successful and Unsuccessful University Students," *Contemporary Educational Technojogy*, No. 1, pp. 36-45, 2010
12. A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, "Clustering for Improving Educational Process Mining," in *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, pp. 11-15, New York, USA, 2014
13. C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting Students' Final Performance from Participation in On-Line Discussion Forums," *Computers & Education*, Vol. 68, pp. 458-472, 2013
14. R. S. Baker, D. Lindrum, M. J. Lindrum, and D. Perkowski, "Analyzing Early At-Risk Factors in Higher Education E-Learning Courses," in *Proceedings of the 8th International Conference on Educational Data Mining*, pp. 150-155, National University for Distance Education, Madrid, Spain, 2015
15. J. Bainbridge, J. Melitski, A. Zahradnik, E. Lauría, S. Jayaprakash, and J. Baron, "Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education," *Journal of Public Affairs Education*, Vol. 21, No. 2, pp. 247-262, 2015
16. J. P. Campbell, "Utilizing Student Data within the Course Management System to Determine Undergraduate Student Academic Success: An Exploratory Study," Doctoral Dissertation, Purdue University, pp. 31-61, 2007

17. J. Bravo, S. Sosnovsky, and A. Ortigosa, "Detecting Symptoms of Low Performance using Prediction Rules," in *Proceedings of the 2nd Educational Data Mining Conference*, pp. 31-40, Universidad de Cordoba, Cordoba, Spain, 2009
18. L. P. Macfadyen and S. Dawson, "Mining LMS Data to Develop an Early Warning System for Educators: A Proof of Concept," *Computers & Education*, Vol. 54, No. 2, pp. 588-599, 2010
19. B. B. Minaei and W. Punch, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System," in *Proceedings of Genetic and Evolutionary Computational Conference*, pp. 2252-2263, Chicago, Illinois, USA, 2003
20. L. V. Morris, S. Wu, and C. Finnegan, "Predicting Retention in Online General Education Courses," *The American Journal of Distance Education*, Vol. 19, No. 1, pp. 23-36, 2005
21. S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, "Predicting Students' Academic Performance using Multiple Linear Regression and Principal Component Analysis," *Journal of Information Processing*, Vol. 26, pp. 170-176, 2018
22. D. Detoni, C. Cechinel, R. A. Matsumura, and D. F. Brauner, "Learning to Identify At-Risk Students in Distance Education using Interaction Counts," *Revista de Informática Teórica e Aplicada*, Vol. 23, No. 2, pp. 124-140, 2016
23. G. Geraldine, M. Colm, O. Philip, and H. Markus, "Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education," *Journal of Learning Analytics*, Vol. 3, No. 2, pp. 330-372, 2016
24. R. Barber and M. Sharkey, "Course Correction: Using Analytics to Predict Course Success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 259-262, Vancouver, British Columbia, Canada, 2012
25. G. D. Chen, C. C. Liu, K. L. Ou, and B. J. Liu, "Discovering Decision Knowledge from Web Log Portfolio for Managing Classroom Processes by Applying Decision Tree and Data Cube Technology," *Journal of Educational Computing Research*, Vol. 23, No. 3, pp. 305-332, 2000
26. Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee, "Targeting the Right Students using Data Mining," in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*, pp. 457-464, Boston, Massachusetts, USA, 2000
27. C. Mi, X. Peng, and Q. Deng, "An Artificial Neural Network Approach to Student Study Failure Risk Early Warning Prediction based on TensorFlow," *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Vol. 219, pp. 326-333, 2018
28. T. Y. Yang, C. G. Brinton, W. C. Joe, and M. Chiang, "Behavior-based Grade Prediction for MOOCs via Time Series Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 5, pp. 716-728, 2017
29. K. A. Capao, A. D. Cantara, A. M. Ceniza, P. M. J. Eduardo, S. B. Polinar, and J. M. Tero, "Predicting Academic Performance with Intelligence, Study Habits and Motivation Factors using Naive Bayes Algorithm," *International Journal of Engineering Research & Technology*, Vol. 5, No. 3, pp. 182-185, 2016
30. M. Tripathi and A. K. Agarwal, "Probabilistic Determination of Student Performance using Naive Bayes Classification Algorithm," *International Journal of Engineering Science and Computing*, Vol. 7, No. 8, pp. 14749-14752, 2017
31. C. Mollica and L. Petrella, "Bayesian Binary Quantile Regression for the Analysis of Bachelor-to-Master Transition," *Journal of Applied Statistics*, Vol. 44, No. 15, pp. 2791-2812, 2017
32. A. K. Hamoud, A. M. Humadi, W. A. Awadh, and A. S. Hashim, "Students' Success Prediction based on Bayes Algorithms," *International Journal of Computer Applications*, Vol. 178, No. 7, pp. 6-12, 2017
33. H. Martin, Z. Zdenek, and Z. Jaroslav, "Ouroboros: Early Identification of At-Risk Students without Models based on Legacy Data," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 6-15, Vancouver, Canada, 2017
34. C. Kevin and A. David, "Utilizing Student Activity Patterns to Predict Performance," *International Journal of Educational Technology in Higher Education*, Vol. 14, No. 1, pp. 1-15, 2017
35. S. M. Jayaprakash, E. W. Moody, E. J. M. Laurí a, J. R. Regan, and J. D. Baron, "Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative," *Journal of Learning Analytics*, Vol. 1, No. 1, pp. 6-47, 2014
36. C. Antunes, "Acquiring Background Knowledge for Intelligent Tutoring Systems," in *Proceedings of the International Conference on Educational Data Mining*, pp. 18-27, Montreal, Québec, Canada, 2008
37. S. Rana and R. Garg, "Evaluation of Students' Performance of an Institute using Clustering Algorithms," *International Journal of Applied Engineering Research*, Vol. 11, No. 5, pp. 3605-3609, 2016
38. M. Kumar, S. Shambhu, and P. Aggarwal, "Recognition of Slow Learners using Classification Data Mining Techniques," *Imperial Journal of Interdisciplinary Research*, Vol. 2, No. 12, pp. 741-747, 2016
39. M. D. Pistilli and K. E. Arnold, "Purdue Signals: Mining Real-Time Academic Data to Enhance Student Success," *About Campus*, Vol. 15, No. 3, pp. 22-24, 2010
40. K. E. Arnold and M. D. Pistilli, "Course Signals at Purdue: Using Learning Analytics to Increase Student Success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 267-270, Vancouver, Canada, 2012
41. A. Essa and H. Ayad, "Improving Student Success using Predictive Models and Data Visualisations," *Research in Learning Technology*, No. 20, pp. 58-70, 2012
42. Z. Zhang, W. Liu, and Z. Han, "Learning Dashboard: A Novel Learning Support Tool in the Big Data Era," *Modern Distance Education Research*, No. 3, pp. 100-107, 2014
43. L. Wang, Y. Ye, and X. Yang, "Design of Online Learning Early-Warning Model based on Big Data," *Modern Educational Technology*, Vol. 26, No. 7, pp. 5-11, 2016
44. P. Kamal and S. Ahuja, "A Review on Prediction of Academic Performance of Students At-Risk using Data Mining Techniques," *Journal on Today's Ideas - Tomorrow's Technologies*, Vol. 5, No. 1, pp. 30-39, 2017
45. Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," in *Proceedings of the Informing Science & IT Education Conferences*, pp. 647-665, Cassino, Italy, 2010

46. C. Romero and S. Ventura, "Data Mining in Education," *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, Vol. 3, No. 1, pp. 12-27, 2013
47. C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. M. Fardoun, and S. Ventura, "Early Dropout Prediction using Data Mining: A Case Study with High School Students," *Expert Systems*: *The Journal of Knowledge Engineering*, Vol. 33, No. 1, pp. 107-124, 2016
48. J. B. Heppen and S. B. Therriault, "Developing Early Warning Systems to Identify Potential High School Dropouts," National High School Center, American Institutes for Research, pp. 1-13, 2008

Chunqiao Mi received his Ph.D. from the College of Information and Electrical Engineering at China Agricultural University in 2012. At present, he is an associate professor in the School of Computer Science and Engineering at Huaihua University, and his research interests include data science and educational information technology.