Review of Academic Early Warning Methods Based on Data Mining

Xingyu Xue[1], Baoqing Xu[1*], Xingyue Wu[1], Leixiao Li[2,3]

[1]*Inner Mongolia University of Technology, School of Information, Hohhot, 010000, China*
[2.]*Inner Mongolia University of Technology, School of Data Science and Application, Hohhot, 010000, China*
[3.]*Inner Mongolia Software Service Engineering Technology Research Center Based on Big data, Hohot, 100080,China*
[*]*Corresponding author: bqxu@163.com*

Abstract: *The main purpose of academic warning is to promptly detect students' learning crises during the teaching process, and to provide effective warning and intervention, thereby achieving supervision of the entire learning process by teachers and students and effectively improving the quality of students' learning. The author used the method of data comparison and analysis to conduct a detailed comparative analysis of data collection and source processing methods, algorithm evaluation methods, and warning result analysis methods in academic warning methods. Then, an analysis was conducted on machine learning, linear regression, artificial neural networks, and improved algorithms based on data mining algorithms, and it was found that there are still some shortcomings in academic warning, such as insufficient universality of warning models; The sample types are generally numerical and textual, and factors such as learners' emotions, learning dynamics, family economy, and background are not included in the scope of data collection. Through the analysis of data mining methods in academic warning, the author proposes that these algorithms have their own advantages and disadvantages, and their respective strengths can be used to complement each other and solve specific problems together.*

Keywords: *Education Big data;Academic Warning;Data Mining;Machine learning;Deep Learning;*

Introduction

Predictive models are increasingly being adopted as a competitive strategy in education. They enable predictive management of students' final grades and teachers' teaching progress while offering early preventive interventions for students at risk of academic failure. An increasing number of studies use predictive models to address learning challenges, underscoring the growing popularity of academic early warning systems.

This study aims to provide a comprehensive and systematic review of the current progress, trends, unresolved issues, and future research directions related to academic early warning systems. The paper begins with a brief introduction to the current research challenges in academic early warnings, followed by a comparison of commonly used models and modeling methods both domestically and internationally. It concludes by summarizing their strengths and weaknesses and exploring future directions for research in the field of academic early warning systems.

1. Research Content of Academic Early Warning

1.1 Research Issues

Academic early warning systems can be divided into three stages: early-term warning, mid-term warning, and end-of-term warning. A critical component is predicting and signaling performance in individual courses. Using data mining technologies, academic early warning systems can accurately identify students who may fail and face grade accumulation issues, possibly leading to repetition or dropout. This predictive capability overcomes the limitations of traditional academic management and achieves a proactive warning effect.

Currently, many universities have established their own academic early warning systems and have achieved certain successes. However, significant challenges remain in the implementation process. For example:

The data used to build academic early warning models is often limited and lacks diversity.

Existing systems are often unable to effectively integrate students, teachers, and administrators.

There are challenges in accurately and promptly predicting students' academic performance.

Timely warnings to students and teachers, along with effective guidance for measures to address risks, remain difficult to implement.

Supporting students and teachers in adopting appropriate strategies to improve study habits and teaching methods, thereby enhancing the overall quality of learning, is still a focal research area.

Addressing these challenges is essential for advancing the effectiveness and adoption of academic early warning systems.

1.2 Research Methods

1.2.1 Evaluation Methods for Algorithms

As a comprehensive system, academic early warning primarily consists of two components: data processing and data regression. In the data processing phase, four key metrics are commonly used for evaluation: accuracy, F-measure, recall, and precision.

1.2.1.1 Accuracy

Accuracy is a metric used to evaluate classification models. A higher value indicates better accuracy. In Equation (1), the variables are defined as follows:

TP (True Positive): Correctly predicted positive cases.

TN (True Negative): Correctly predicted negative cases.

FP (False Positive): Incorrectly predicted positive cases.

FN (False Negative): Incorrectly predicted negative cases.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad （1）$$

1.2.1.2 F-measure

The F1 score, also known as the F1-measure or F1-Score, is a metric for evaluating classification problems. It is the harmonic mean of precision and recall. The F1 score ranges from 0 to 1, where a value closer to 1 indicates better model performance, as shown in Equation (2).

$$F1 = 2*P*R(P+R) \qquad （2）$$

1.2.1.3Recall

Recall represents the proportion of actual positive samples that are correctly predicted as positive by the model. It is often used to evaluate the model's coverage or completeness in identifying positive cases, as shown in Equation (3).

$$\text{Recall} = \frac{TP}{TP+FN} \qquad （3）$$

1.2.1.4 Precision

Precision refers to the proportion of samples predicted as positive by the model that are actually positive. It is commonly used to evaluate the model's accuracy in identifying positive cases, as shown in Equation (4).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad （4）$$

In data regression, the Pearson correlation coefficient and root mean squared error (RMSE) are commonly used metrics.

a. Pearson Correlation Coefficient: The Pearson correlation coefficient between two variables is defined as the ratio of their covariance to the product of their standard deviations, as shown in Equation

(5).

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma x \sigma y} \qquad (5)$$

b. Root Mean Squared Error (RMSE): RMSE is a widely used metric for regression problems. It represents the average error between the predicted values and the actual values. In the formula, y_i represents the actual value, $y^\wedge{}_\mathrm{i}$ represents the predicted value, and nnn denotes the sample size, as shown in Equation (6).

$$\mathrm{RMSE}{=}\sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - \mathrm{y}^\wedge{}_\mathrm{i})\dot{2}} \qquad (6)$$

1.2.3 Analysis of Early Warning Results

In terms of early warning performance, Afzaal et al. [1], Albalooshi [2], and Chen & Cui [3] analyzed learners' classroom performance to predict their exam scores. Zeineddine et al. [4] predicted first-semester grades based on freshmen's admission performance and entrance scores and used cumulative metrics to predict learners' final grades while evaluating their graduation outcomes.

In addition to predicting grades, early warning systems can identify at-risk students, provide assistance, and implement intervention plans tailored to their needs. For example, Rafique [5] and Zacharis [6] predicted which students are most likely to fail in the early stages. Research by Gray and Perkins [7] and Chui [8] identified students at risk of dropping out, while Kabathova & Drlik [9] determined course withdrawal risks based on academic performance.

Table 1 The Accuracy of Algorithms in Academic Early Warning Systems

method	Algorithm name	average accuracy（%）
classification	decision tree（J48）	82.0
	Random Forest	96.1
	Support Vector Machine	96.0
	Naive Bayes	85.0
Regression algorithm	logistic regression	92.6
	linear regression	96.2
cluster	K-means clustering	93.2
Association Rules	Class association rules	88.2

In the evaluation of classification algorithms, the random forest model by Kabathova and Drlik [10] achieved high precision, with an accuracy rate of 86%, recall rate of 96%, F1 score of 91%, and an overall accuracy of 93% in predicting students' course withdrawal levels.

In another study, Lincke et al. [11] identified the gradient boosting tree and XGBoost as the best models for predicting the probability of students answering quizzes correctly, with accuracy rates of approximately 88% and AUC values of 0.903 and 0.94, respectively. Since the ROC curve is more suitable for balanced datasets, and given the imbalanced distribution of categories in the dataset, the authors used an alternative metric called Precision-Recall to evaluate their models.

In regression function evaluation, Jensen et al. trained a random forest regression (RF) model to predict students' exam scores, achieving a Pearson correlation coefficient of 0.53, indicating relatively high accuracy.

2. Data Mining-Based Algorithms

2.1 Machine Learning-Based Data Mining Algorithms

2.1.1Decision Tree

The decision tree algorithm is a significant classification algorithm in machine learning. Its fundamental idea is to use an existing dataset for training to generate a tree-like model. This model can then be used to predict and classify new data. The advantage of decision trees is that they enumerate all feasible solutions to the decision problem, along with various possible natural states, and calculate the expected values of each feasible solution under different states. Decision trees visually represent the entire decision-making problem across different stages in terms of time and decision sequence. When applied to complex multi-stage decisions, they provide clear stages and hierarchical structures, facilitating group discussions by decision-making bodies. This allows for comprehensive consideration of various factors, aiding in making accurate decisions.

2.1.2 Artificial Neural Networks (ANN)

Artificial Neural Networks are network structures capable of solving practical problems with multiple nodes and multiple output points. Their advantage lies in their ability to rapidly find optimized solutions. Solving optimization problems often requires significant computational resources. Using a feedback-based ANN specifically designed for a problem can leverage the high-speed computation capabilities of computers, enabling the rapid discovery of optimal solutions.

2.1.3 Random Forest

The fundamental principle of the random forest algorithm is to use the Bootstrap self-sampling method to obtain different sample sets for model construction, thereby increasing diversity among models and improving the capability for extrapolated predictions. The algorithm has three key advantages:

Versatility: It can address both classification and regression problems, handling categorical and numerical features simultaneously.

Resistance to Overfitting: By averaging decision trees, it reduces the risk of overfitting.

Stability: Random forests are highly robust. Even if a new data point is introduced into the dataset, it has minimal impact on the overall algorithm. It may affect a single decision tree but is unlikely to influence all the trees in the forest, ensuring consistent performance.

2.1.4 Naive Bayes

The Naive Bayes algorithm assumes that the categories in academic early warning data are mutually independent. It calculates the posterior probability of each sample belonging to each class using Bayes' theorem and then assigns the sample to the class with the highest posterior probability. This method is particularly useful in cases where simplifying assumptions about data independence can lead to efficient classification.

Table 2 Comparison of input data for commonly used algorithms

Algorithm name	key column	key column	predictable column

Decision trees and random forest algorithm	must include a numerical or textual column to uniquely identify each record. Composite keys are not allowed.	One or more input columns, which can be discrete or continuoc input attributes will affect processing time	requiring at least one predictable column. Multiple predictable attributes can be included in the model, and the types of these predictable attributes can be different, either numerical or discrete. However, increasing the number of predictable attributes will result in longer processing times.
Cluster analysis algorithms	must include a numerical or textual column to uniquely identify each record. Composite keys are not allowed.	Each model must contain at least one input column that contains the values used to generate this classification. You can have as many input columns as needed, but it depends on the number of values in each column. Adding additional columns will increase the time required to train the model.	This algorithm does not require predictable columns to generate the model, but can add predictable columns of almost any data type. The values of predictable columns can be considered as inputs to the clustering analysis model, or they can be specified for prediction purposes only.
Naive Bayes algorithm	requires each model to include a numerical or textual column that uniquely identifies each record. Composite keys are not allowed	All input types must be discretized data or data that has undergone discretization processing must have at least one predictable column.	it must contain discrete or discretized values. The values of predictable columns can be used as inputs and are often processed as inputs. In order to find the relationship between each column
neural network algorithms	one	One or more	One or more

2.2 Data Mining Algorithms Based on Artificial Neural Networks

Artificial Neural Networks (ANNs) play a vital role in predicting learning outcomes. For instance, Neha et al. [12] proposed a deep neural network to predict students' grades. Dias et al. [13] and Mubarak et al. [14] utilized Long Short-Term Memory (LSTM), a type of recurrent neural network, to predict course grades, evaluate teacher-student interaction effectiveness, and assess classroom engagement as well as weekly student performance.

Deep learning introduces a method where computers automatically learn pattern features and integrate feature learning into model building, reducing the incompleteness caused by manual feature design. Certain applications based on deep learning in machine learning have surpassed existing algorithms in recognition or classification performance under specific conditions. However, in scenarios with limited data, deep learning algorithms cannot provide unbiased estimates of data patterns. Achieving high accuracy requires extensive data support. Additionally, the increased complexity of graphical models in deep learning significantly raises algorithm time complexity. Ensuring real-time performance requires advanced parallel programming skills and superior hardware support. Consequently, only well-funded research institutions or enterprises can use deep learning for cutting-edge, practical applications.

2.3 Data Mining Algorithms and Their Improved Variants

Madan integrated score-based association rule techniques with decision tree algorithms, including ID3, C4.5, and CART. They improved the original class-ratio-based information requirement in the ID3 algorithm by redefining attribute divisions, reducing model runtime to 10% of its original duration. Zhou Hangyu [15] refined academic warning systems by predicting scores at the granularity of knowledge points and question types. For course withdrawal warnings, they combined XGBoost and LightGBM algorithms, increasing accuracy from 78.27% to 82.57%. This model achieved high prediction accuracy and improved robustness.

Meng Jiao combined association rule algorithms with random forest algorithms to identify relationships between courses, enhancing the model's applicability. Pradeep used three decision tree and rule induction algorithms in WEKA. Decision trees were converted into a set of IF-THEN rules, aiding teachers in guiding struggling students to avoid future failures. Rule induction allowed the expansion of derived rules, providing satisfactory descriptions for each category.

2.4 Advantages and Limitations of Major Algorithms

In the data preprocessing phase, common techniques include clustering algorithms (e.g., K-means, hierarchical clustering, density-based clustering) and association rule mining algorithms. Clustering algorithms are limited to numerical features, while association rule mining algorithms explore relationships between students' academic data. For instance, Apriori, a common association rule mining algorithm, sets support and confidence thresholds to generate rules. However, the generated rules may not always align with the relationships between courses in academic data.[16]

In the modeling phase, neural network algorithms excel at predicting overall course scores but struggle to provide predictions for individual chapters. Xu et al. analyzed the correlation between internet usage behavior and scores, finding a strong relationship. However, data extraction and analysis rely heavily on deep packet inspection (DPI) technology, requiring interdisciplinary collaboration. XGBoost demonstrates strong data fitting capabilities but has numerous parameters, and their selection affects prediction accuracy. Decision tree algorithms have low time complexity (logarithmic with respect to data points) and are interpretable as white-box models. However, they are prone to overfitting. Logistic regression models offer low computational cost and storage requirements but are prone to underfitting and have limited accuracy. While deep learning algorithms have gained popularity in recent years, their application in academic early warning remains limited.

From the perspective of early warning effectiveness, current results often have coarse granularity, offering only binary classifications (e.g., pass/fail) with limited interpretability. Students cannot use the results to make targeted improvements in their studies. Early warnings are typically limited to an alert without providing detailed analyses of why students fail to complete their studies. Moreover, educational administrators often lack insights into the factors contributing to students' academic struggles.

Overall, academic early warning systems span multiple disciplines and fields. However, the lack of collaboration among researchers from different disciplines remains a significant barrier to addressing the aforementioned issues.

3 Summary and Outlook

This paper conducted a comprehensive review and analysis of academic early warning systems in terms of data preprocessing, system modeling, and model evaluation.

First, student learning data, behavioral trajectories, and device usage patterns can be collected from school academic affairs offices, personnel departments, network centers, and telecom companies. These data can be analyzed as parameters to evaluate students' habits, learning methods, and attitudes. Second, association rule mining algorithms can be considered for data preprocessing, while random forest, J48, and Naive Bayes algorithms can be utilized for modeling. Among these, random forest offers robust performance overall, J48 excels in data conversion processes, and Naive Bayes performs well in data classification. Third, model evaluation should use four key metrics: accuracy, F1 score, recall, and precision. Finally, algorithms should be validated across different majors, courses, grades, and instructors to comprehensively assess their effectiveness.

Academic early warning systems enable teachers to intervene in students' learning processes promptly, adopting tailored management strategies and teaching methods to achieve personalized learning and guidance. These systems help students recognize their academic standing, identify gaps with peers, anticipate potential outcomes, and adjust their learning methods and attitudes proactively.

This paper compared and analyzed the latest research on data mining in academic early warning systems, highlighting the advantages and limitations of major algorithms. Researchers can refer to these insights based on their specific objectives. Furthermore, as big data and algorithms evolve, future research in academic early warning can integrate data from graphics, audio, and visual sources to enhance accuracy, reliability, practicality, and intelligence.

Fund Projects

2022 Youth Science and Technology Talent Project of Higher Education Institutions (NJYT22084); National Natural Science Foundation Project: "Research on Artificial Emotion Generation Mechanisms for Mongolian Dialogue Systems" (62166029); Basic Research Fund Project for Universities Directly Under the Autonomous Region: "Research on Artificial Emotion Generation Models for Dialogue Systems" (JY20220074); Key Teaching Research Project of Inner Mongolia University of Technology (020421001).

References

[1] Afzaal M, et al.Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation[J]. Frontiers in Artificial Intelligence,2021, 4(2):1-5.

[2] Albalooshi F, AlObaidy, H., & Ghanim, A. Mining students outcomes: An empirical study[J]. International Journal of Computing and Digital Systems,2019, 8(3), 229–241.

[3] Chen F & Cui Y. Utilizing student time series behaviour in learning management systems for early prediction of course performance[J]. Journal of Learning Analytics,2020,7(2), 1–17.

[4] Zeineddine,Braendle U, & Farah A.Enhancing prediction of student success: Automated machine learning approach[J]. Computers & Electrical Engineering, 2020,89:106903.

[5] Rafique A,et al.Integrating learning analytics and collaborative learning for improving student's academic performance[J]. IEEE Access,2021,9:167812–167826.

[6] Zacharis N Z.Classification and regression trees (CART) for predictive modeling in blended learning[J].International Journal of Intelligent Systems and Applications,2018,10(3):1–9.

[7] Gray C C & Perkins D. Utilizing early engagement and machine learning to predict student outcomes[J].Computers and Education,2019,131: 22–32.

[8] Chui K T, et al.Predicting at-risk university students in a virtual learning environment via a machine learning algorithm[J].Computers in Human Behavior,2020,107:105584.

[9] Kabathova J & Drlik M.Towards predicting student's dropout in university courses using different machine learning techniques[J].Applied Sciences (Switzerland), 2021,11(7):3130.

[10] Hussain M, et al. Student engagement predictions in an e-learning system and their impact on student course assessment scores[J].Computational Intelligence and Neuroscience, 2018,10(15):47-63.

[11] V Heilala, M. Saarela, P. Jääskelä and T. Kärkkäinen, Course Satisfaction in Engineering Education Through the Lens of Student Agency Analytics[C]//IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 2020, pp. 1-9.

[12] Hew K F, et al. What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach[J],Computers & Education,2020,145:103724.

[13] Lincke A, et al. The performance of some machine learning approaches and a rich context model in student answer prediction[J].Research and Practice in Technology Enhanced Learning,16(1).

[14] Khan I, et al. An artificial intelligence approach to monitor student performance and devise preventive measures[J].Smart Learning Environments,2021, 8(1).

[15] Yang S J, et al. Predicting student's academic performance using multiple linear regression and principal component analysis[J]. Journal of Information Processing,2018,26:170–176.

[16] Omer U, Farooq M S & Abid A.Cognitive learning analytics using assessment data and concept map: A framework-based approach for sustainability of programming courses[J]. Sustainability (Switzerland), 2020,12(17):69-90.