Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080204
ISSN 2371-8838 Vol. 8 Num. 2

Academic Behavior Analysis and Early Warning System Based on K-Means Algorithm

Feiyang Huang[1,a,*], Renteng Li[1,b], Shuyu Shi[1,c], Chong Zhang[1,d]

[1]*College of Artificial Intelligence, North China University of Science and Technology, Tangshan, China*

[a]*nai_wei_lan@163.com*, [b]*18132504795@163.com*, [c]*16613903870@163.com*, [d]*15383705130@163.com*

[*]*Corresponding author*

Keywords: K-Means Algorithm, Multiple Linear Regression, Academic Warning, Grade Prediction, Django

Abstract: This paper presents the design and implementation of an academic behavior analysis and warning system based on the K-Means algorithm. The system combines students' historical grades with current behavioral data to construct a predictive and academic warning model, aimed at assisting educators in quickly identifying academic risks and providing adjustment suggestions for students on the academic edge. The system is divided into registration and login modules, administrator modules, and user modules, realizing functions such as identity authentication, permission allocation, and account management. In the model construction phase, K-Means clustering is applied to training samples, and multiple linear regression models for grade prediction are built based on the clustering results. In the testing phase, grades of experimental groups are predicted and error analysis is conducted. Experimental results show that the system has lower prediction errors in the construction of predictive models for intelligent medical engineering majors and higher prediction accuracy for computer science and technology majors. The system also establishes a four-level warning mechanism, represented by red, orange, yellow, and green, to help users intuitively understand their academic situations. Overall, this study provides effective support for student academic development through a K-Means-based grade prediction system, with practical application value.

1. Introduction

Against the backdrop of continuous reform in educational mechanisms and increasingly fierce educational competition, the analysis of student performance and related behaviors has become an important topic in educational research. With the widespread adoption of big data analysis and intelligent algorithms, many scholars have conducted extensive research on academic warning issues, such as Huang Fangliang et al.'s data mining algorithms [1] and Su Jin's fuzzy association algorithms [2]. This article aims to study the feasibility and practical effects of using the K-Means algorithm combined with students' historical grades and current behavioral data for prediction and academic warning. By delving into the intrinsic connections between student performance and

learning behavior, we hope to construct an efficient predictive model that can identify students who may face academic challenges in advance and provide targeted adjustment suggestions.

2. Introduction to Related Algorithms

2.1. K-Means Clustering Algorithm

The K-Means clustering algorithm is an unsupervised learning method that divides data samples into different clusters based on their similarities. The similarity between data samples in space is generally measured using Euclidean distance. The algorithm is widely used due to its ease of interpretation, simplicity, and good scalability to data [3]. The core idea of the algorithm is to randomly select k samples as initial cluster centers C_k ($1 \le i \le k$, i is an integer) in the space of data samples, then calculate the similarity between each data sample x and the cluster center C_k, assign the samples that are more similar to the cluster center C_k to the same cluster, update the cluster center, and iterate until the cluster center no longer changes or reaches the maximum number of iterations. The calculation formula is as follows:

$$d(x, C_k) = \sqrt{\sum_{i=1}^{n} (x_i - C_{ki})^2} \tag{1}$$

where x is a data sample; C_k is the k-th cluster center; n is the data dimension; x_i and C_k are the i-th attribute values of the data sample and the cluster center C_k, respectively.

2.2. Multiple Linear Regression

Multiple linear regression algorithm is a classic and mature algorithm, and its mathematical model is as follows: Let the dependent variable be Y, and the independent variables be X_1, X_2, ... , X_P, there exists the following linear relationship:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon \tag{2}$$

where β_0 is the regression constant; β_1, β_2,..., β_P are the regression parameters for each independent variable X_1, X_2, ... , X_P; ε is the error term, which follows a certain distribution $\varepsilon \sim N(0, \sigma^2)$.

3. Grade prediction based on the K-Means algorithm

3.1. Calculation Process

The prediction process of student grades based on the K-Means algorithm and multiple linear regression can be divided into two main stages: model establishment and model solving.

In the model establishment stage:

Step 1. Perform K-Means clustering on the training samples.

Step 2. Divide the training samples based on the clustering results.

Step 3. Construct multiple linear regression models for grade prediction for different categories of training samples.

In the model testing stage:

Step 1. Experimentally predict the grades of the experimental group using the constructed multiple linear regression grade prediction models.

Step 2. Analyze the prediction results for errors.

3.2. Evaluation Metrics

For the model establishment described above, the evaluation metric used is Mean Absolute Error (*MAE*). It is defined as the average of the absolute deviations between individual observed values and true values. Its advantage is that it avoids errors canceling each other out and accurately reflects the magnitude of actual prediction errors. The formula for calculating *MAE* is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|Score_p - score\right|}{100} \tag{3}$$

where N represents the number of data samples, $Score_p$ and $Score$ represent the predicted score and the original score, respectively. A smaller *MAE* value indicates smaller prediction errors and higher prediction accuracy of the model.

3.3. Experimental Data

The experimental data selected for this study involved extensive statistical experiments on the actual grades of students majoring in Intelligent Medical Engineering and Computer Science and Technology at a certain university. The Computer Science and Technology major included grades from 60 students in three sophomore courses and their graduation grades, while the Intelligent Medical Engineering major included grades from 57 students in ten freshman courses.

To protect the privacy of students, this study concealed the students' identity information such as names, gender, and student ID numbers in the original data, while retaining their grades in each subject. Box plots were used to preprocess the data, removing students with grades greater than $\bar{X}+3\sigma$ and less than $\bar{X}-3\sigma$. Finally, 55 Computer Science and Technology students and 53 Intelligent Medical Engineering students were selected.

For the evaluation of student grades, considering the differences in majors and courses, the study first used the entropy weighting method to calculate the weights of various indicators for both majors. Seven common and relatively weighted indicators were selected: course audiovisual learning, homework completion, attendance, average score in online classroom tests, study time outside of class, weighted score from the previous semester, and GPA score.

3.4. Display of Some Student Grades

Partial course student grades for the Intelligent Medical Engineering major are shown in Table 1.

Table 1: Partial Course Student Grade Table for Intelligent Medical Engineering Major

Name	Homework Completion (Unit: %)	Attendance (Unit: %)	Average Score in Online Classroom Tests (as a percentage of the total score)	Final Grade
1	100	100	84	95
11	100	100	89	91
21	100	100	95	89
31	90	100	86	80
41	90	97.5	84	74
51	85	97.5	83	65

Partial course student grades for the Computer Science and Technology major are shown in Table 2.

Table 2: Partial Course Student Grade Table for Computer Science and Technology Major

Name	Homework Completion (Unit: %)	Attendance (Unit: %)	Average Score in Online Classroom Tests (as a percentage of the total score)	Final Grade
1	100	100	95	95
11	100	100	93	94
21	100	100	89	92
31	10	100	88	89
41	90	97.5	85	80
51	80	97.5	89	75

4. Experimental Results and Analysis

The experiment conducted in this study consisted of six sets of comparative experiments on the data from the two majors. One course grade from each major was randomly selected as the training sample, and the remaining course grades were used as experimental samples for predicting results. The training samples were clustered using K-Means, and the training samples were divided into different categories based on the clustering results. Considering the number of training samples, two sets of experiments were conducted with k=2 and k=3, which means dividing the training samples into two and three categories, respectively, and constructing multiple linear regression grade prediction models for each category. The MAE value was recorded for each experiment, and the average MAE value was taken as the final performance evaluation metric. The experimental results are shown as follows Figure 1:

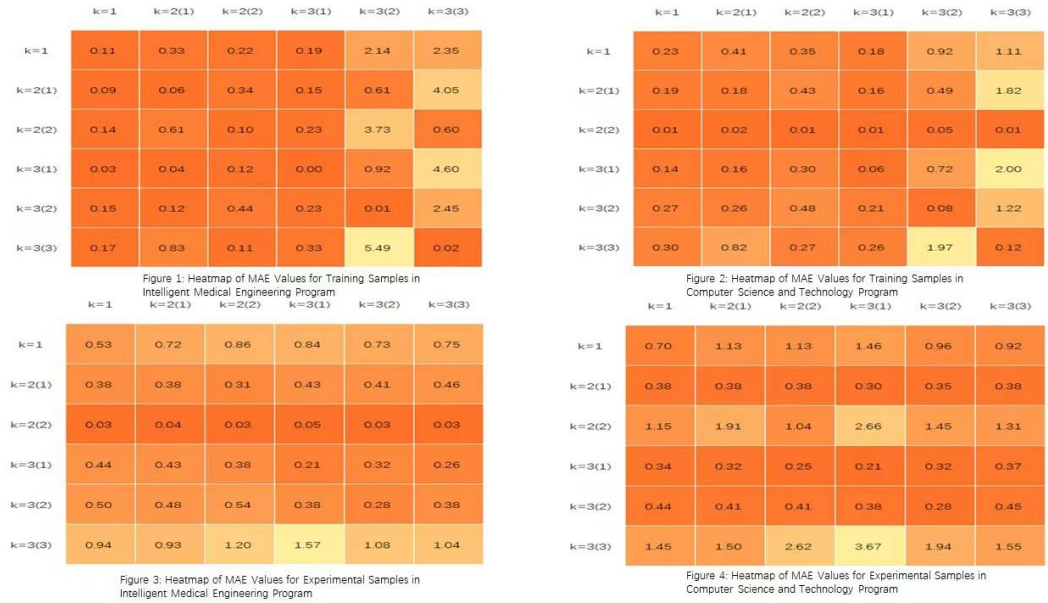

Figure 1: Experimental Results Diagram

From the experimental results of the two majors mentioned above, it can be seen that the prediction model for the Intelligent Medical Engineering major constructed with k=2 has a lower prediction error. Moreover, its prediction accuracy in the Computer Science and Technology major is also significantly higher than that of the other experimental groups. Therefore, the final grade prediction model is as follows:

$$Y = -45.825 - 0.121X_1 + 0.018X_2 - 0.205X_3 + 0.136X_4 - 0.022X_5 + 0.061X_6 + 41.143X_7 \quad (4)$$

In the equation, Y represents the predicted score, X_1 represents the course audiovisual learning, X_2 represents the homework completion, X_3 represents the attendance, X_4 represents the average score in online classroom tests, X_5 represents the study time outside of class, X_6 represents the weighted score from the previous semester, and X_7 represents the GPA score of the student.

The predicted results obtained from the experimental group using the above multiple linear regression equation are shown as follows Table 3 and Table 4.

Table 3: Partial Grades of Intelligent Medical Engineering Students

Name	Actual Grade	Predicted Grade
1	95	94.66
11	91	90.57
21	89	90.34
31	80	79.68
41	74	76.01
51	65	64.58

Table 4: Partial Grades of Computer Science and Technology Students

Name	Actual Grade	Predicted Grade
1	95	94.89
11	94	95.24
21	92	90.89
31	89	88.68
41	80	81.24
51	75	74.91

5. Introduction to Related Technologies

5.1. Introduction to the Django Framework

Django is an advanced Python web framework based on the MVC (Model-View-Controller) design pattern. It provides a carefully designed set of tools that significantly simplifies the development process of web applications. The Django framework includes various built-in features such as user authentication, URL routing management, template engine, and object-relational mapping (ORM). These features allow developers to focus more on implementing business logic without delving into the underlying technical details[4]. By using Django, developers can build web applications with higher efficiency, stronger security, and better maintainability to meet diverse business needs. Additionally, Django has shown significant advantages and wide applications in the development of websites for small and medium-sized enterprises. Numerous real-world cases demonstrate that the combination of Django framework and Python language can effectively support the design and implementation of various applications, such as the radiotherapy process management system[5] and web calculator[6].

5.2. Introduction to MySQL Database

MySQL is a relational database management system that holds a significant position in the current technology field and is widely popular. The system follows the standard SQL (Structured Query Language) specification, allowing it to run on multiple operating systems and be compatible

with various programming languages such as Python, C, and C++. MySQL database is favored by developers for its compact size, efficient data processing speed, reliable data exchange mechanism, and excellent concurrency control features. Moreover, the widespread use of the MySQL database indicates its good scalability and maintainability, providing convenience for future technological upgrades and feature expansions.

6. Overall Function Design and Implementation

6.1. Design of MySQL Data Tables

In the system studied in this paper, several data tables were specially designed to support the requirements of academic behavior analysis. These data tables include AdmInfo (Administrator Information Table), StuInfo (Student Information Table), and StuProgram (Student Behavior Information Table), among others. Each data table defines corresponding fields according to actual requirements. The specific structure and attributes of these fields can be found in Tables 5-7 of this paper.

By designing these data tables reasonably and combining them with the efficient performance of MySQL database, this system can achieve rapid storage, querying, and analysis of academic behavior data. This provides a solid foundation for subsequent academic research and practical applications.

Table 5: AdmInfo

Name	Type	Description
id	bigint	Primary Key, Auto Increment
uid	varchar	Account
password	varchar	Password

Table 6: StuInfo

Name	Type	Description
uid	varchar	Primary Key, Student ID
name	varchar	Name
password	varchar	Password

Table 7: StuProgram

Name	Type	Description
id	bigint	Primary Key, Auto Increment
subject	varchar	Subject
video	varchar	Course audiovisual learning
work	varchar	Homework completion
signin	varchar	Attendance
test	varchar	Average score in online
time	smallint	Study time outside of class
score	decimal	Weighted score from the previous semester
gpa	decimal	GPA score
uid_id	varchar	Foreign Key, associated with uid in StuInfo

6.2. Design of Basic Functional Modules

The system presentation end integrates HTML5, CSS, and other technologies, and uses Bootstrap as the front-end template to ensure consistent page design style, with good human-computer interaction. The analysis and warning system is divided into three modules: registration and login module, administrator module, and user module, as shown in the figure 2 below:

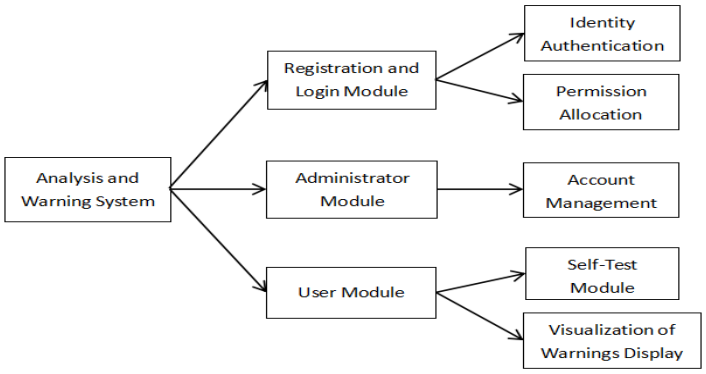

Figure 2: System Module Display

6.2.1 Registration and Login Module

In the user authentication system designed in this paper, users need to accurately enter their account and password on the specified login interface when attempting to log in to their account. If any of the account or password entered by the user is incorrect, the system will immediately display an error message below the corresponding input box to clearly inform the user that their login attempt was unsuccessful. This immediate feedback mechanism helps users quickly identify input errors and correct them. Conversely, if the account and password entered by the user are correct, the system will automatically perform a redirect operation to guide the user to their corresponding user module homepage. This design aims to enhance user experience, reduce unnecessary operational steps, and make the login process more efficient.

This module also applies to administrator login. Administrators can enter their specific account and password through the same login interface and process as regular users. Once verified, the system will automatically guide them to the administrator module homepage. This design ensures the consistency and convenience of the administrator login process, effectively improving the overall usability of the system.

When a user has not created an account, the system allows them to complete the account creation process by clicking the registration button. The account, name, and password information registered by the user will be saved to the StuInfo database using the statement StuInfo.objects.create(uid=uid, name=name, password=password).

6.2.2 Administrator Module

The administrator module is used to save and display user account, name, and password information, and provide the function to delete specified accounts.

6.2.3 User Module

After successfully logging in, users can enter the self-assessment module through the

hyperlinked text in the navigation bar. In this module, users can freely fill in the subject name they want to predict, and provide their own learning behavior information on the course mobile end. After receiving the information, the system backend uses the final grade prediction model mentioned above to obtain the predicted grade result for the corresponding subject, and transmits the result to the front-end page.

It is worth mentioning that this system has built a four-level warning mechanism to help users intuitively understand their situation: Red warning indicates that the predicted grade for the subject is below the passing line, the situation is serious, and the learning status needs to be adjusted urgently; Orange warning indicates that the predicted grade for the subject is in the range of 60 to 65 points, there is a certain risk of failure, adjustments should be made to avoid further deterioration; Yellow warning indicates that the predicted grade for the subject is in the range of 65 to 70 points, which requires the user's attention; Green warning indicates that the predicted grade for the subject is above 70 points, the situation is good and can be maintained. The specific warning module is shown in the figure 3 below:

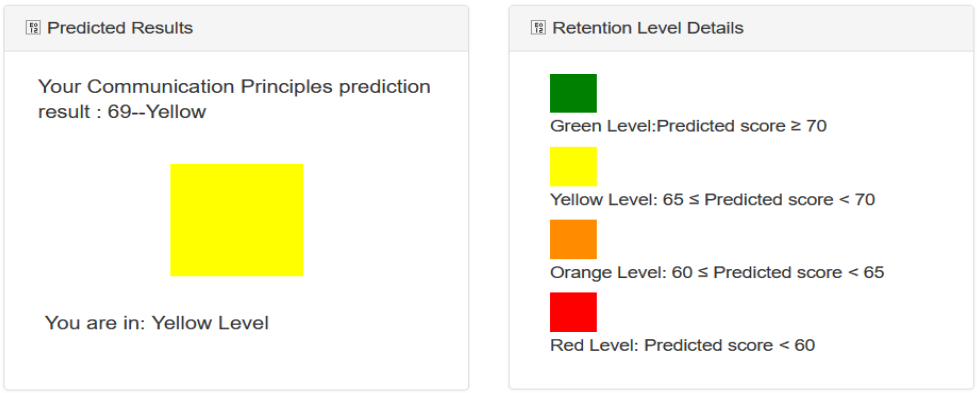

Figure 3: Specific Warning Module Diagram

7. Conclusion

As individual entities, students' academic performance is closely related to their learning behavior. This paper uses the K-Means algorithm combined with students' historical grades and learning behavior to predict and analyze their academic information, effectively realizing the transformation from "results" to "warnings". Through this research, it is hoped that targeted adjustment suggestions can be provided for students on the academic margin.

References

[1] Huang Fangliang, Xu Huanqing, Shen Tongping, Jin Li, Yu Lei. Design and Experimental Research on Intelligent Learning Effect Early Warning Management System Based on Data Mining. Journal of Tonghua Normal University, 2022, 43(12): 84-89

[2] Su Jin. Student Academic Early Warning Method Based on Fuzzy Association Rules. Software, 2022, 43(7): 27-29

[3] Wang Guanbang, Liu Hongyan, Li Jinsong, et al. Research on Student Performance Prediction Method Based on K-Means. Information Technology, 2023, 47(02): 1-6. DOI:10.13274/j.cnki.hdzj.2023.02.001.

[4] Guo Henan. Website Design and Implementation Based on Django and Python Technology. Digital Communication World, 2023(6): 60-62

[5] Fu Tingyan, Sun Yaping, Liang Rui, Wu Xiaodong, Fan Ting, Zhao Wangxiong, Zhang Ke. Design and Implementation of Radiotherapy Process Management System Based on Django. Modern Hospital, 2023, 23(5): 750-754

[6] Wang Yufen, Zhao Dandan. Design and Implementation of Web Calculator Based on Bootstrap and Django Framework. Information and Computer, 2023, 35(1): 143-146