2022 2nd International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR) | 979-8-3503-3426-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICBAR58199.2022.00032

Student Behavior Analysis Research of based on Data Mining

Shaofen Fan[1*], Yanyan Xu[2], Baolin Zhu[2] and Lei Chen[3]
[1]JiNan University,Guangzhou, Guangdong province, 510632,China
[2] East China Normal University,Shanghai, 200241,China
[3] NanJing Audit University,Nanjing, Jiangsu province,211815, China
[*]Corresponding author's e-mail: fanshaofen@jnu.edu.cn

Abstract. **University data governance produces a large amount of data, including a large number of student data. How to use data mining technology to analyze and apply student data is the key to improve the scientific management level of the school and the quality of talent training. This paper proposes an adaptive K-means clustering algorithm. Using data mining and big data technology, it selects student consumption, learning, book borrowing, access record and other characteristic data, and conducts data collection, processing and clustering analysis of student behavior data by establishing a characteristic model of student behavior data. The experiment shows that the algorithm used in this paper analyzes the student feature model, and the students who eat regularly and go to the library more often generally have better academic performance. The results of data analysis can be used to predict student performance. Based on the data mining technology, this paper analyzes the data of students' behavior to realize the prediction and early warning of students' behavior, which can provide data decision and scientific support for the precise management of schools.**

Keywords- Big Data, adaptive K-means cluster algorithm, Cluster analysis, Student behavior, Data governance

I. INTRODUCTION

With the rapid development of Internet, artificial functions, big data, cloud computing and other technologies, lots of universities have constructed smart campus with many information systems. The information system generates a large amount of data. More and more universities have fully realized the importance of data. Big data is characterized by massive, high-speed, diversity, and value. Data has become the most important asset to ensure the sustainable and healthy development of university information. President Xi said that we should use big data to improve the modernization of national governance. The effective use of educational data is an important basis for universities to realize information innovation. It's the cornerstone for promoting the deep reform of teaching, scientific research, management and service methods, and realizing the all-round modernization of higher education[1]. At present, each department of the university has its own information system, and has accumulated a large number of student data, including consumption information, access control, book borrowing, academic performance and other data. Under the new era and new situation of rapid development of big data, how to speed up the unified data platform, use various big data and data mining technologies to mine the potential value of data from student behavior data is the key to improve the scientific management level and talent training quality of the school.[2] Through the analysis of student behavior data, we can provide decision-making support for college education and teaching, and establish a mechanism of "speaking with data, making decisions with data, managing with data, and innovating with data".

II. RELATED WORK

The International Data Management Association(DAMA) believes that data governance is the exercise of rights and control, including planning, monitoring and implementation, in the process of data asset value creation, so as to effectively manage data assets and continuously create data value [3]. China's Education Modernization 2035 proposes to carry out education governance capacity optimization action supported by big data, and promote the whole process of education and teaching by means of information technology such as the Internet with the help of data mining [4]. After building a data center, the school will gather a large amount of data, including a large number of student behavior data. The traditional student management mostly relies on subjective judgment and experience, which makes it impossible to analyze students objectively, leading to subjective management decisions and lack of scientific basis and data support. Machine learning clustering algorithms are widely used in student behavior analysis [5], but in traditional clustering algorithms, the mean is very sensitive to outliers, which often affect the final clustering analysis results. Therefore, this paper proposes a K-means clustering algorithm based on the elbow method, which uses adaptive K value selection to improve the accuracy and accuracy of clustering.

III. DATA GOVERNANCE PLATFORM CONSTRUCTION

Data is the basis for the modernization of education governance. Through data governance, college data can become standard, authoritative, available, easy to use, accurate, unified, shared and interconnected. The effective application of data governance can make university management more intelligent and agile, decision-making more scientific and management more efficient. The university data governance platform can include: data acquisition layer, data governance layer, data bazaar layer, and data service layer.

- Data acquisition layer

The business data and dynamic real-time data accumulated in the history of each information system form the data source, and the multi-source and heterogeneous data in the data source are stored in a physical database in full 1:1, forming the data lake (ODS). This data collection method not only solves the problems of cross database access, but also ensures the data

979-8-3503-3426-5/22/$31.00 ©2022 IEEE
DOI 10.1109/ICBAR58199.2022.00032

security of the data source. Any subsequent processing of the data in the data lake will not affect the data source of the original business system even if data loss occurs in the data governance process. At the same time, the data center is relatively independent from each information system through the intermediate database of the data lake, as shown in Figure 1. If the business system is replaced by a supplier in the future , the new manufacturer only needs to provide the data API interface according to the college standard, and the data center only need to adjust the database connection information to complete the connection with the business system, without necessity for secondary integration development.

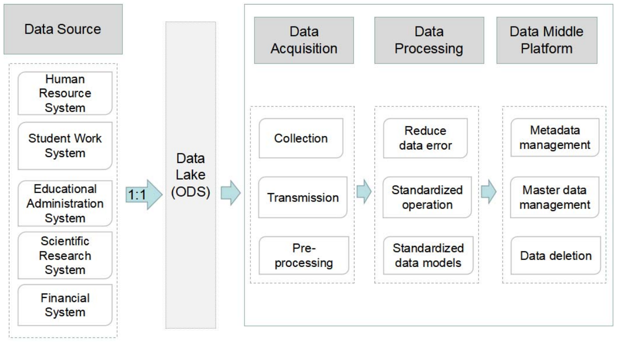

Figure 1. Schematic Diagram of Data Governance

- Data governance layer

The data governance layer is the process of forming a standardized data warehouse after the original data in the data lake is processed through authoritative source identification, data de duplication, format conversion, multi table association and other steps to clean, govern, integrate and transform. The standardized data warehouse, namely the master data center, is the most important data asset in the college's data governance and serves as the only authoritative standard data resource at the college level.

- Data mart layer

The standardized data warehouse provides various data services externally through the data mart, such as student data sets, teacher data sets, scientific research data sets and so on. And it’s available for various departments to apply for use in the form of applicable interfaces, such as API, ETL and download files. The data mart includes three typical forms: subject repository, thematic repository and indicator repository[7].

- Data application layer

The data application layer is the goal and attribution of data governance, and the achievement of data governance construction. Through big data, data visualization, report and other technologies, it can provide data decision support for school education and teaching, and serve the development of the college. Its application forms include “one table”, personal portraits of teachers and students, leadership drive, etc.

The college data governance project has built the data center recently, developed a unified data standard for the college, built an authoritative master data platform, realized the collection, governance and sharing of data from various departments, formed an authoritative data asset of college, and realized the interconnection and sharing of data in the college's information systems.

IV. Student Behavior Analysis Research

Based on the data governance platform of the school, the system can obtain all aspects of students' data from student engineering, all-in-one card, library, access control and other systems. Undergraduate student data is shown in Figure 2.Through comprehensive analysis of various data of students, it is helpful for colleges and counselors to fully understand students' performance in school in a timely manner, so as to better provide personalized guidance and help for students. Through the analysis of various kinds of data of students, it can also be used as the auxiliary data for the evaluation of students and provide data support for the accurate decision-making of the school. For example, students can be depicted and analyzed by various data such as their all-in-one card consumption, daily access to the school gate, weekly access to the library, and wireless Internet access duration, so as to understand students' preferences in the activity area, daily work and rest, and to warn against potential safety hazards. Reflect whether the students are really poor from the consumption information of all-in-one card, and provide data decision support for the selection of poor students[8].

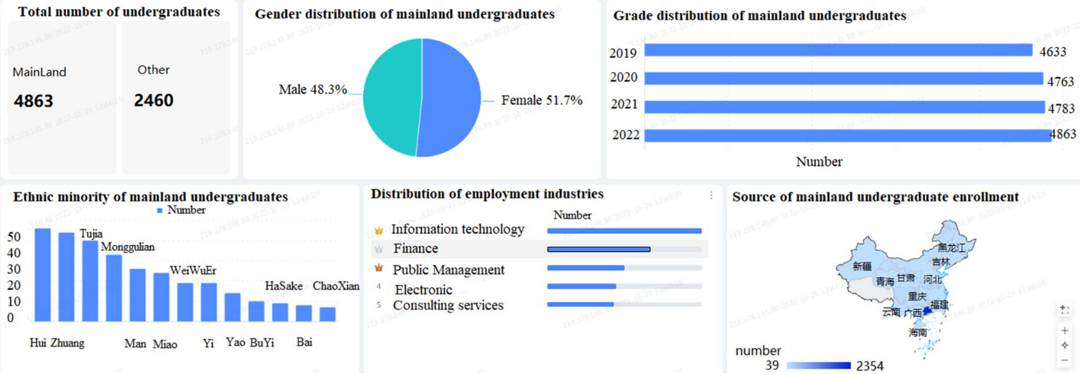

Figure 2. Undergraduate student data diagram

A. *Student data feature modeling*

First, we need to obtain the students' original data from the data center. The original data will have a lot of noise, so we need to preprocess the students' original data and delete some irrelevant noise data. Then we conduct feature modeling on the data, mainly extracting students' basic information, all-in-one card consumption data, access card records, and libraries as feature attributes. The architecture of student behavior analysis is shown figure 3.

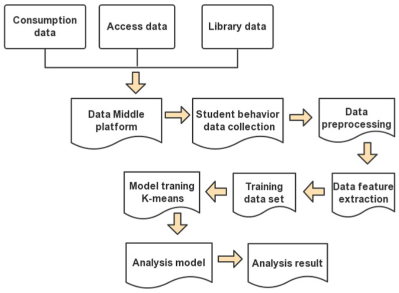

Figure 3. Architecture diagram of student behavior analysis

(1) Basic information of students: This information describes the basic information of students, such as name, student number, college, major, telephone number, grade, etc.

(2) All in one card information: mainly including the student's average monthly consumption amount (avgMonConsumption), average consumption amount per time (avgConsumption), and the monthly dining rate (DiningRate) and other characteristic information. Before data analysis, pay attention to pre-processing the data, such as deleting the consumption data of graduated students and the consumption data during the holidays, to eliminate some unstable effects of the holidays and improve the accuracy of clustering and decision-making.

$$\text{avgConsumption} = \frac{\text{Co}\mathit{nsumption}}{\mathit{time}} \times 100\% \tag{1}$$

$$\text{avgMonConsumption} = \frac{\text{Co}\mathit{nsumption}}{\text{Month}} \times 100\% \tag{2}$$

(3) Access card records: including information about students entering and leaving the campus gate and the study room.

(4) Library information: including the records of books borrowed by students, the time and times of entering and leaving the library every month, etc.

Extracting data for association analysis is usually divided into two steps. First, we need to calculate the support degree of each candidate item set, and filter according to the minimum support degree and minimum confidence degree preset by the algorithm to obtain frequent item sets. Support is the probability of occurrence, that is, the proportion of occurrences of the current item in all item sets. Confidence is a conditional probability, that is, the probability of condition Y when condition X occurs. The calculation of confidence is as follows:

$$\text{SUPPORT}(\text{X} \rightarrow \text{Y}) = \text{P}(\text{X} \cup \text{Y}) \tag{3}$$

$$\text{Confidence}(\text{x} \rightarrow \text{y}) = \text{p}(\text{Y}|\,\text{X}) \tag{4}$$

B. *Methods of student behavior analysis*

(1) Normalization

As there are many characteristic attributes involved in student behavior data, it is necessary to standardize all kinds of characteristic attribute data. We map the characteristic data with different value ranges to the [0,1] interval (as shown in the formula below[9]), so as to eliminate the impact of different value ranges on the results. X_{normal} is the standardized data, X is the original data, and $\text{X}_{\min}$ and $\text{X}_{\max}$ are the maximum and minimum values of the original data set.

$$\text{X}_{normal} = \frac{\text{X} - X_{\min}}{X_{\max} - X_{\min}} \tag{5}$$

(2) K-Means clustering algorithm

K-Means clustering algorithm is an iterative clustering analysis algorithm[9]. Its central idea is: divide the data into K groups, randomly select K objects as the initial clustering center, calculate the distance between each object and each seed clustering center, and assign each object to the nearest clustering center. Each time a sample is allocated, the cluster center of the cluster will be recalculated according to the existing objects in the cluster. This process will be repeated until a termination condition is met.

Assume that the initial data set D={x_1, x_2, ... , x_n}, x_i, x_j is the data object, x_{il} is the l characteristic attribute of x_i, and x_{jl} is the l characteristic attribute of x_j. Definition of The Euclidean distance formula is shown as below[10]:

$$\text{d}(\text{x}_\text{i}, \text{x}_\text{j}) = \sqrt{\sum_{\text{l}=1}^{\text{n}} (\text{x}_{\text{il}} - \text{x}_{\text{jl}})^2} \tag{6}$$

The quality of K value selection has a great impact on the analysis results of K-means clustering algorithm. In this paper, K-Means elbow method is used, that is, K value is selected by calculating SSE (sum of squares of errors)[11]. The specific idea is: SSE describes the compactness of each cluster sample to a certain extent; Therefore, the smaller the SSE, the better the clustering effect.

$$\text{SSE} = \sum_{\text{i}=1}^{K} \sum_{x \in C_i} (C_i - x)^2 \tag{7}$$

C. *Experiment and Results of Student Behavior Analysis*

The clustering K value is set as 4, and the iteration termination threshold is 0.01. The Euclidean distance is used as the distance between data objects, and the K-means clustering algorithm is used for analysis. The clustering analysis results of student consumption are shown in Figure 4. According to the analysis results, students can be divided into four categories according to their consumption level: high, high, medium and low. The results of cluster analysis can be combined with the actual situation of students to help judge the economic situation of students' families[11]. Compared with the usual subjective judgment of students' economic situation, the poor students are graded and the scholarship is evaluated. The cluster analysis based on data mining provides data support and decision support for the school management.

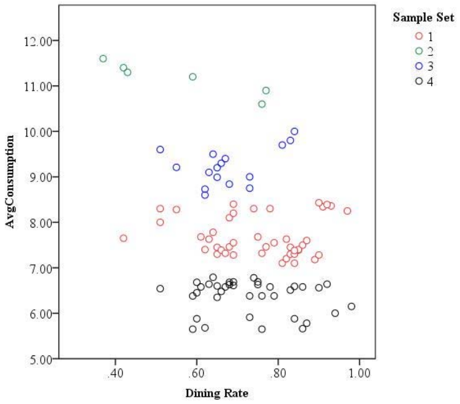

Figure 4. Cluster analysis results based on all-in-one card consumption

Cluster Analysis and Classification Results of Student Behavior Data is shown in Table1. Students are divided into more, more, average and less according to the times of borrowing books and entering and leaving the library[12]. We can judge whether students are in the school from the records of students entering and leaving the school gate. This data can be used as an auxiliary means of student management. Counselors can use the data to predict in advance, so as to better serve and manage students.

Table 1. Cluster Analysis Results of Student Behavior

Clusters		Set 1	Set 2	Set 3	Set 4
All in one card System	Avg Consumption	7.68	11.17	9.23	6.37
	Dining Rate	0.75	0.56	0.68	0.73
Library System	Book Borrowing	27	5	14	20
	Reading Time	40	7	28	15
Access System	Access to library	58	16	47	36

Through comprehensive analysis of students' reading time in the library every month, the number of times in and out of the library and students' daily consumption records, students' learning habits and achievements are predicted. According to daily experience, students who spend a long time reading in the library and have regular daily work and rest have better academic performance, while students who usually have irregular daily work and rest and seldom go to the library to read have poor academic performance. Compared with the final grades of students in the actual database, the results predicted by cluster analysis(Figure 5)have good consistency, which shows that the cluster analysis method in this paper is effective and has a good predictive role.

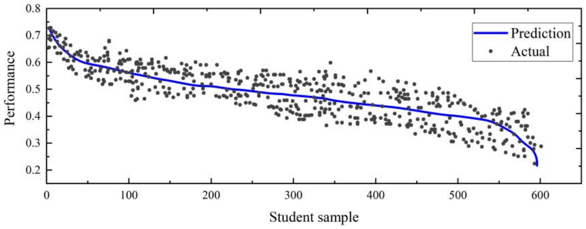

Figure 5. Comparison of the results predicted by cluster analysis and actual

V. Conclusion

This paper proposes an adaptive K-Means clustering algorithm, which uses data mining technology to analyze and mine student behavior data, and through the analysis of various characteristic data of students, it realizes the prediction of student performance. The experiment shows that the adaptive K-Means algorithm adopted in this paper can classify, count and analyze student behavior data, thus providing decision and data support for school management. For example, K-means clustering algorithm is used to cluster the student consumption data, and the real poor students are obtained by clustering the daily consumption data of students. Based on the actual situation of our school's data governance project, this paper analyzes the student behavior data through data mining technology, provides diversified data services for teachers and students, and boosts the modernization of school governance. Informatization has always been on the way, and promoting the modernization of school governance and the high-quality development of education and teaching through data governance will certainly be a cause that universities need to make long-term efforts and adhere to.

ACKNOWLEDGMENTS

This paper is one of the phased achievements of Jiangsu University's philosophy and social science project "Research on oriented Audit Informatization Talent Training Mode" (2022SJYB0337).

REFERENCES

[1] Peng Y and Li Y. (2020) Research on higher education data ecological governance system from the perspective of smart campus. China Educational Technology, vol 5:88-100.

[2] Han X.(2021) SPSS Analysis of Daily Consumption of Poor College Students from the Perspective of Field Theory. 2021 International Conference on Education, Information Management and Service Science,pp:377-380.

[3] DAMA International, The DAMA guide to the data management body of knowledge. Technics Publications, New Jersey.

[4] XiaoYing L and Shouwu H.(2021) Research and Analysis of Student Portrait Based on Campus Big Data. 2021 IEEE the 6th International Conference on Big Data Analytics, pp:23-27.

[5] Zhou W. (2021) Big Data-driven Decision-making in Universities: Patterns, Problems and Optimization Strategies. Research on Educational Development, vol 9:78-84.

[6] Xinhua News Agency, 2019. The Central Committee of the Communist Party of China and the State Council issued China Education Modernization 2035. http://www.gov.cn/zhengce/2019-02/23/content_5367987.htm

[7] Qin YY and Liao HY.(2020) Bottleneck and countermeasure analysis of smart campus data governance construction. Modernization of Education, vol 7:100-101.

[8] Liu ZD. (2020)Practice Research on University Data Center Construction from the Perspective of Data Governance. ITCA, pp:489-492.

[9] Mao ZJ, Wu JY, Qiao YL and Yao H. (2021).Government data governance framework based on a data middle platform. Aslib Journal of Information Management, vol 74, No.2, pp:289-310.

[10] Qin XG and Xue Y, (2022)Research on the governance framework and ecological system construction of university education data. Higher Education Forum, No.2, pp:23-28.

[11] Jiang GWX, (2021)Practical Exploration and Theoretical Discussion on University Affairs Data Governance--Taking Peking University as an Example.Beijing Education(Higher Education), No.4, pp:8-11.

[12] Liu YQ, Mao WH, Wu C, Yu JQ and Wang SX, (2022) Promote Data Governance Practice fro University Teachers Based on the “One Table” Platform. Modern Educational Technology, Vol 32, No.1, pp:118-126.