Design of Student Portrait Model Based on Educational Big Data Mining

Wenwen Yin[1,a],Jiatong Ju[2,b],Youquan Gao[1,c],Kun Fan[1,d],Jinkai Chen[1,e] ,Zhaohui Zheng[1,f*]

Email:[a]yinchusheng@gmail.com, [b]1985874163@qq.com ,[c]3561374641@qq.com, [d]3079248130@qq.com,[e]2054289394@qq.com,[f]zhengzhaohui@wit.edu.cn

[1]Wuhan Institute of Technology, School of Mathematics and Physics, Wuhan, Hubei, China,
[2]Wuhan Institute of Technology, School of Computer Science & Engineering, Wuhan, Hubei, China

Abstract. The correlation between student behavior and academic development is a key focus of school education work in the context of big data. This article designs a diversified and integrated education big data association mining model. Firstly, the behavioral data of middle school students in supermarkets, canteens, psychology, and education are collected and synthesized into a data table according to their student IDs. Then, student labels are extracted based on the scale standards and partition functions. Then, based on the FP growth association algorithm, the degree of association and differences in behavior performance, consumption level, and academic level among different student groups are studied, Finally, the tree hole text and logistic regression model were used to construct student portraits, predict psychology, and academic trends, respectively. The experimental results indicate that the constructed student portrait can effectively describe students' academic and life characteristics, providing a basis for educators to provide personalized care and support to students to a certain extent.

Keywords: Association mining, Student portraits, Educational Data Mining

1 Introduction

Educational big data[[1]] refers to large-scale data collections collected, stored and analyzed in the field of education. It covers a variety of data from students, teachers, schools, and education systems, which are stored in large-scale databases and analyzed and mined to provide deep understanding of educational processes and outcomes. The education big data care model utilizes the massive data generated in the teaching process to capture educational characteristics, analyze the reasons, and help educational managers better teach students in accordance with their aptitude.

With the rapid development of data mining technology, more and more educational researchers hope to extract the knowledge and information behind educational data. For example, Yu Minghua[[2]] studied the construction of student portraits based on visualization technology, and presented personal portraits for students and group portraits for teachers with the help of data visualization technology; Yang Menghui[[3]] used association rules to mine the potential relationship between multi-dimensional features and graduation destinations; Zhang Tao[[4]] proposed a Design a general framework for a learner model to achieve a sharable and reorganizable learner data model; Tang Qian[[5]] uses education data to mine student portraits and con-

ICEKIM 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2337321

duct learning tests to provide educators with an in-depth diagnosis of students' learning status and provide precise teaching guidance and intervention for reference.

Khan[[6]] proposed an educational big data model for improving management processes and making informed decisions, improving the effectiveness of educational managers' decision-making; Beatrice Bonami[[7]] proposed a multi-modal process to better understand the relationship between education and new technology prospects Contact; Anna Y Q H[[8]] The study collected data sets from three universities and found that the factors that affect the prediction performance of classification methods are the number of features and the number of categories.

However, the current educational big data research mainly focuses on the existing portrait research focusing on single behavioral characteristics or the improvement of educational data algorithms, but there is a lack of data-driven models to mine the characteristics of group caring education. Therefore, this paper proposes a student portrait and care model based on educational big data mining. Through accurate student portraits and personalized support, it can meet the needs of students in learning, emotion and social interaction, and promote their learning growth, mental health and social development.

2 Data collection and preprocessing

2.1 Data collection

This study takes 803 senior high school students from a middle school in Wuhan as an example, including four types of data: basic information data, academic data, life data, and psychological data. The above data is obtained through API. Basic information includes family members, political status, height, weight, etc; Due to the limited scope of student activities in school, such as rest and consumption places, life data can be obtained through the student card.

The data here includes three meal frequency characteristics, three meal consumption average characteristics, supermarket consumption average characteristics, and dormitory rest time characteristics; Academic data is extracted from the performance management system; The psychological data is sourced from a psychological questionnaire SCL-90[[9]] that students fill out regularly. In order to simplify subsequent data analysis, the four types of data were integrated into one dataset. The student data information is shown in the Table 1 below：

Table 1. Basic information of the research data in this article

	Field Information	data sources
Essential information	Student ID, gender, political status, nature of household registration, height, weight, family status	Xuexin System
Life data	Average consumption of three meals, average consumption of three meals, average consumption of supermarkets, frequency of supermarket consumption, rest behavior, bathing behavior	One Card System

Academic data	Monthly exam results, mid-term results, final results, classroom performance	Academic Affairs System, Online Camera
Psychological data	Anxiety level, depression level, physical discomfort, sleep and diet level, interpersonal relationships, vent tree holes	Psychological Care System

2.2 Data procession

Due to the large amount of repetitive data such as characters and time involved in daily life data, there is a large amount of empty data in academic data, and basic information and psychological data are constantly changing. So Python is used to exclude outlier, remove duplicate values and save all data in an Excel table. Due to the requirements of the new curriculum reform to weaken students' attention to academic performance and ensure their privacy, this article transforms academic data. Based on the difficulty coefficient of each exam and combined with actual teaching experience, a fuzzy score partition function can be constructed. By incorporating the score data into the partition function, students' grades can be fuzzily mapped to real numbers between 0 and 1. The definition of the fuzzy partition function is:

$$\mu(x) = \frac{x_i - P\_min}{P\_max - P\_min}, P\_min \leq x \leq P\_max \tag{1}$$

The value range is $\mu(x)$Is the degree of membership, P_max is the maximum score, P_min is the minimum score. This article sets the 0-0.2 level to be improved, the 0.2-0.8 level to be good, and the 0.8-1 level to be excellent.

3 Label System and Association Rule Mining

3.1 Label construction

The core of student portrait construction is to have labels that describe student characteristics, so constructing student labels is the foundation of all work[[10]]. The basic information of students is directly labeled according to their categories, while life data and psychological data are divided according to the Student Health Handbook and SCL-90 scale standards. This article uses a partition function to obtain three levels of academic data: excel, good, and to be improved. Table 2 shows the student portrait label system:

Table 2. Student Portrait Label

Label and meanings		Labels and meanings		Labels and meanings		Labels and meanings	
A1	Graduating in 2023	E 1	Serve as class committee	I2	Moderate consumption frequency	…	
A2	Graduating in 2024	E2	Not serve as a class committee member	I3	Low consumption frequency	L1-1	No depression

A3	Graduating in 2023	F1	Male	J1	Living habits	L1-2	Slightly depressed
ID	Student mark	F2	Female	J2	Irregular life	L1-3	Severe depression
B1	Rural	G1	High level of consumption	K1-1	Good overall score	L2- 2	No anxiety
B2	City	G2	Moderate level of consumption	K1-2	Good overall	L2- 2	Slightly anxious
C1	CPC member	G3	Low level of consumption	K1-3	Overall performance needs to be improved	L2- 2	Severe anxiety
C2	Masses	H1	Reasonable length of rest	K2-1	Excellent in mathematics	L3-1	Excellent interpersonal relationship
D1	Meals	H2	Rest time needs to be improved	K2-2	Good at math	L3-2	Good interpersonal relationship
D2	Regular meals	H3	Insufficient rest time	K2-3	Mathematics scores need to be improved	L3-2	Poor interpersonal relationship
D3	Irregular meals	I1	High consumption frequency	K3-1	Excellent in Chinese		

3.2 Association Mining Algorithm

The analysis of educational data in this article adopts association mining algorithms[[11]], which can find the correlation rules between students' behavior, academic performance, and psychological habits, helping teachers and schools understand students' interests and needs. The implementation process of association mining is as follows: Firstly, scan the data table based on the student ID to generate frequent 1_Itemset, then arranged in descending order of support, again based on frequency 1_ Scan the data table for itemsets and construct an itemset association FP tree. Continue to recursively search for all frequent itemsets on the FP tree, and finally find all association rules in the frequent itemsets based on the support and confidence of the set.

3.3 Association analysis of group mining

Use the FP-Growth algorithm to mine the association rules between the characteristics of the three types of student groups (excellent, good, and to be improved), identify the association rule table between various characteristics, and set the minimum support threshold and minimum confidence The thresholds are set to 0.3 and 0.5, and the frequent pattern trees of the three types of student groups are obtained as shown in Figure 1.

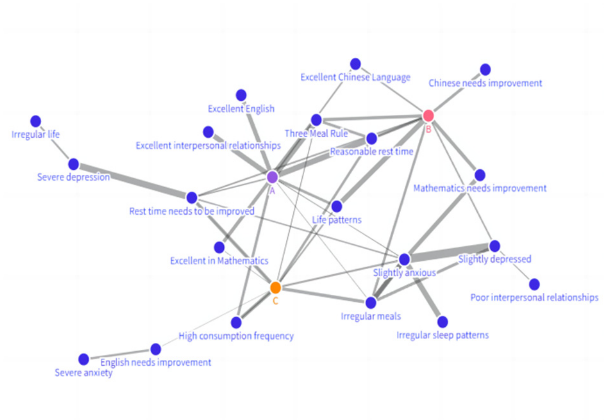

Fig. 1. group student association map

In addition, based on the calculations of support, confidence, and improvement, due to the complexity of the association graph, this article only extracted the first two strong association rules from the frequent pattern trees of the three types of student groups mentioned above, as shown in Table 3.

Table 3. Association rule calculation table

Grade	Frequent Itemset	Support	Confidence	Improvement
A	{ Regular meals , Reasonable rest time }→A	0.3256	0.5801	1.123
	{ Excellent interpersonal relationships }→A	0.3855	0.5743	1.143
B	{ Chinese needs improvement }→B	0.3960	0.5156	1.213
	{ Slightly anxious, math needs improvement }→B	0.5727	0.7161	1.146
C	{ Irregular life, Need to improve rest time, Severe depression }→C	0.4212	0.7070	1.129
	{ Irregular meals, Poor interpersonal relationships, Slight depression }→C	0.3402	0.4402	1.246

These rules demonstrate the connections between some student features, and the student feature association rules shown in the above table are effective when the support level is greater than 1. According to Table 3, the calculation results show that group A students perform well in various characteristics, have a regular lifestyle, and have good grades; Group B students have weaknesses in certain subjects and may experience mild anxiety; Group C students have irregular schedules and serious academic problems, and they have psychological problems that require strengthened care.

4 Student Portrait

Combining the results of association mining with student tree hole data, this article uses text mining technology based on word segmentation processing to classify and count students' relevant learning and life keywords, obtain high-frequency demand words, and visualize them;

Visualize the demands of Class C students as shown in the figure, and it is found that students have demands such as "I feel sleep", "under pressure", and "take a vacation". Figure 2 shows the demands of Class C students; In each category, for individual students in this category, the psychological evaluation SCL-90 scale data is used to analyze their psychological status. The calculation program performs functional statistics on each question, and the logistic regression algorithm is used to infer changes in students' grades and psychological values. This helps teachers understand students' development status in advance and make targeted behavioral changes.

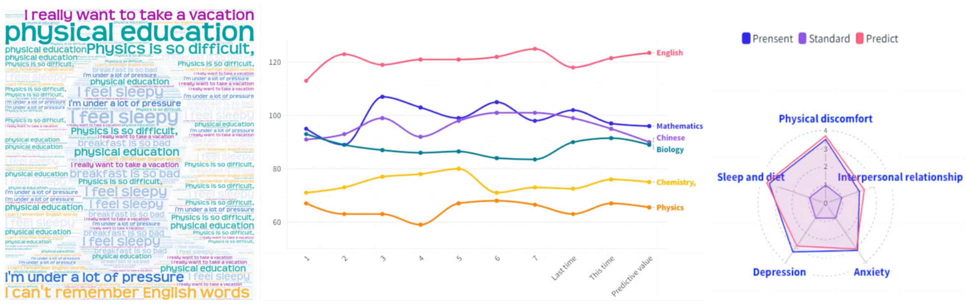

Fig. 2. C group students' demands and analysis and prediction of a certain student's situation

5 Conclusions

This article proposes a multivariate integrated association mining model based on partition function, which aims to address the singularity of education big data mining and the low differentiation of traditional association mining based on precise numerical truncation. The model utilizes partition functions and integrates student psychology and demands for association mining, and takes the performance and psychological data of a certain high school student in Wuhan as an example to analyze and construct a student portrait The results show that the obtained association rules can truly reflect the connections between the academic, life, and psychological aspects of the three types of student groups, providing scientific decision-making and care basis for various teaching activities in middle schools. The later work will focus on the bias and fairness issues of the algorithm, to avoid unfair treatment of certain student groups due to the training of unfair data in the model.

References

[1] Nian-Shing C, Chengjiu Y, Pedro I, Joseph P, et al. Educational Big Data: Extracting Meaning From Data For Smart Education[J], Interactive Learning Environments, 2020, 28(2): 142-147.

[2] Yu Minghua, Zhang Zhi, Zhu Zhiting . Research on the Construction of Research Learning Student Portraits Based on Visual Learning Analysis [J]. China Audio-Technology Education , 2020(12): 36-43.

[3] Yang Menghui, Ren Chao, Xie Baoling, He Zhanxi, Zhang Cairong. Student Employment Portraits Based on Campus Behavior Characteristics [J/OL]. Library Forum: 1-11 [2023-05-21]

[4] Zhang Tao, Zhang Si. Study on the Design and Computation of the Learner Model of Educational Big Data Mining [J]. Electronic Education Research, 2020, 41(09): 61-67. DOI: 10.13811 /j.cnki.eer .2020.09 .009.

[5] Tang Qian. Research on the Construction of Student Portraits and Predictive Analysis of Academic Affair Based on Educational Data Mining [J]. Modern Information Technology, 2023, 7(04): 193-198. DOI: 10.19850/j.cnki.2096-4706.2023 .04.048.

[6] Khan, Mohammed Arshad, and Maysoon Khojah. "Artificial intelligence and big data: The advent of new pedagogy in the adaptive e-learning system in the higher educational institutions of Saudi Arabia." Education Research International 2022 (2022): 1-10.

[7] Beatrice Bonami, Luiz Piazentini, Andre Dala-Possa. Education, Big Data and Artificial Intelligence: Mixed methods in digital platforms[J], Comunicar, 2020, 28(65): 43-51.

[8] Anna Y Q H, Owen H T L, Jeff C H H, C. J Y, Stephen J H Y, et al. Predicting Students' Academic Performance By Using Educational Big Data And Learning Analytics: Evaluation Of Classification Methods And Learning Logs[J], Interactive Learning Environments, 2020, 28(2): 206-230.

[9] Antonio Preti, Mauro Giovanni Carta, Donatella Rita Petretto. Factor structure models of the SCL-90-R: Replicability across community samples of adolescents.[J], Psychiatry Research-neuroimaging, 2019, 272(): 491-498.

[10] Wenbo X, Song S, Hak-Keung L, Jian Z, et al. A Polynomial Membership Function Approach for Stability Analysis of Fuzzy Systems[J], IEEE Transactions on Fuzzy Systems, 2021, 29(8): 2077-2087.

[11] Lior S, Philippe F, Rami Y, Itai D, et al. A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data[J], Information Sciences, 2021, 553: 353-375.