Abnormal Student Detection Model Based on Student Feature Extraction

Integrated Learning Based on Clustering and Neural Network

Shuo Zhang(✉) and Xiangchao Wen

University of Electronic Science and Technology of China, No. 2006, Xiyuan Ave, Chengdu, China
Zs020896@163.com

Abstract. Moral education is an important mission of colleges and universities, and the identification and tracking of abnormal students is an important part of moral education. However, problems such as large amount of data related to college students and the majority of text information are prominent. In this paper, the semantic library is used to quantify the text information, and the emotional feature vector of students is obtained. Combined with the digital data, the feature vector of students is effectively obtained. The K-Means clustering and neural network ensemble learning are used to realize the identification of characteristic students, and the accuracy rate can reach 82.02%, which has certain reference significance.

Keywords: Abnormal students · Semantic Library · K-Means Clustering · Neural Network · Integrated Learning

1 Introduction

In 2018, General Secretary Xi Jinping emphasized the cultivation of socialist construction and successors for the comprehensive development of moral, intellectual, physical, aesthetic and labor at the National Education Conference. In the critical period of socialist modernization construction, the demand for innovative, compound and applied talents is unprecedentedly high, and higher education is also increasingly emphasizing the comprehensiveness. On the one hand, the comprehensiveness is the comprehensive improvement of students' comprehensive quality, and on the other hand, it is the success of students in all aspects, that is, one cannot be less, and one cannot out of line.

However, it is inevitable that some college students will cause problems in various aspects due to family, academic, emotional and other factors. These students are also characterized as abnormal students, which requires the focus, careful guidance and careful guidance of college ideological and political staff. The sudden outbreak of new corona pneumonia has brought great impact on higher education. In the post-epidemic era, universities have also carried out normalized epidemic prevention and control, and students cannot return to their living conditions before the epidemic. Many researchers have formed a unified conclusion through questionnaires, talking and theoretical analysis [1–3]. The poor effect of online courses leads to academic difficulties, false information such as rumors leads to psychological anxiety, parents and children get along for

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 957–965, 2023.
https://doi.org/10.2991/978-94-6463-034-3_98

a long time leads to tension, long-term online communication leads to self-closure, employment situation tension leads to future confusion and other issues. At present, the types of abnormal students are increasing, the number is soaring and the concealment is enhanced. How to find abnormal students and accurately guide abnormal students has become an ideological and political topic worthy of further study in the post-epidemic era. However, there are many problems in colleges and universities, such as the large amount of student data, the large amount of text data and the difficulty of quantitative identification. At present, the ideological and political staff still rely on counsellors. Psychological center teachers and other ideological and political staff to identify abnormal students through heart-to-heart talk or other student feedback, resulting in the difficulty and lateness of finding abnormal students, and the difficulty of identifying and predicting abnormal students.

In view of the above problems, this paper first calibrates and classifies abnormal students based on the author's student work experience and relevant research and discussion analysis, and then realizes the identification of text databases such as student talk database and heart interview database through semantic database, so as to obtain the emotional feature vector of students. Combined with digital data sets such as student information database and campus database, the feature vector of students is obtained by splicing. With the student feature vector as the input, the recognition and tracking of abnormal students are realized by combining clustering and neural network.

2 Student Data Source Statement

At present, there is no clear definition and definition of abnormal students in colleges and universities. By referring to relevant literature [4] and combining with the author's work experience, this paper calibrates and classifies abnormal students as follows and indicates the data sources used in this paper.

2.1 Calibration and Classification of Abnormal Students

Abnormal students refer to students who have obvious abnormalities compared with other normal students due to their own and external factors, and abnormal students generally have obvious abnormalities in academic performance, speech and behavior.

(1) Students with special physiology. Due to congenital or acquired reasons, students with physical disabilities and major diseases are obvious. Some students are difficult to live independently due to physiological defects, and serious psychological problems such as inferiority, self-abandonment, and psychological distortion will occur.

(2) Family abnormal students. The imperfection of native families leads to the abnormality of students, such as single parent families, divorced families, domestic violence families, parents' disability, parents' illegal and criminal behaviors, left-behind children, orphans, etc. This part of the students due to the imperfect performance of the original family for economic difficulties, personality loneliness, self-closure, cognitive difficulties, abnormal anxiety.

(3) Economic abnormal students. This kind of students are mainly native families can't afford the students' tuition fees, living expenses, mainly including the minimum living security students, filing card students, martyrs' children, disabled children, etc. This part of the students mainly for family and other reasons lead to the students' economic difficulties and can't continue to normal academic and university life.

(4) Academically abnormal students. Normal intelligence but due to various reasons lead to serious hanging, lack of credit lead to students who stay and drop out. The main reasons for this type of students are the depression of will caused by games, the inability to concentrate due to psychological problems, and the inability to follow up the learning progress due to the poor basic education in the place of college entrance examination.

(5) Psychological abnormal students. Because of their own or external and other reasons lead to students' psychological adaptation problems, psychological disorders or mental illness, mainly due to academic and psychological pressure, self-awareness deviation, emotional instability, such as autism, depression, mania, schizophrenia.

(6) Students with abnormal communication. Such students are mainly manifested in two extremes. One extreme is that students' social fear is caused by their own psychological reasons, and then they complete self-closure through the Internet, which is difficult to communicate with people. The other extreme is the extreme self of students, always deny others esteem self, others can't communicate with normal.

2.2 Description of Database

(1) Talk database. Conversation is one of the important magic weapons of ideological and political staff in colleges and universities, especially college counsellors The 43rd Decree of the Ministry of Education clearly puts forward that college counsellors should strive to become students' mentors and friends of healthy life. Conversation has become an important means for counsellors to understand each student's ideological and political situation in a timely and effective manner. For example, the author's unit clearly requires counsellor to submit not less than 400 records of students' conversations in the system every semester, and requires not less than 400 records of contacts with students' parents every semester, and some colleges and universities take paper records of conversations. So talk database can be used to identify and track abnormal students.

(2) Psychological interview database. Colleges and universities are equipped with psychological centers. And freshmen need psychological tests after admission. Psychological test data, psychological center interview records and weekly statements of class psychological members are important databases for special student recognition.

(3) Students' basic information database. Students' information collection and updating are carried out through online questionnaires and system filling before enrollment or at the beginning of each semester, mainly including students' family address, health status of family members, statistics of disaster losses in the past three years, per capita annual income of families, hobbies, awards and other related contents, which can be used as data.

(4) School life database. The relevant data of students in the school can be used as a reference database. Student card consumption data, some colleges and universities have realized the identification of poor students based on student card consumption data; students entering and leaving school card recording data, regular entry and exit students may be abnormal; students during school scholarships awarded, subject competition awards, work-study, etc.
(5) Special student library. Through various ways, counsellors and other ideological and political personnel clearly classify some abnormal students, that is, to master their characteristic labels (such labels are few and the sample distribution is uneven).

A student's big data sink is formed based on the above data set to identify abnormal students.

3 Special Student Recognition Model Based on K-means Clustering and Neural Network Ensemble Learning

In this section, the student's feature vector is obtained by quantifying the student's information, and the data we need to build the model is obtained by combining the student's feature vector with the student's label data. We deal with these data by negative sample enhancement and feature dimension reduction so that our model can achieve better accuracy. Finally, we use K-Means clustering and BP neural network to integrate learning, and achieve higher recognition accuracy than traditional BP neural network.

3.1 Data Pre-processing

Since the student data contain a large number of text data and there are many different types of data, in this section, we first conduct a semantic analysis of the text file to quantify the text file. In order to facilitate the training of the subsequent model, we normalize the data, reduce the dimension and enhance the negative sample after stitching the student features of multiple dimensions.

3.1.1 Semantic Analysis Based on Emotional Dictionary

Based on the author's experience and related research, this paper constructs the emotional dictionary of text recognition for different abnormal students, which is shown in Table 1 to facilitate the text information recognition of abnormal students.

This paper obtains the correlation matrix between students and abnormal students through the emotional analysis of database text information, and uses the data with the highest coherence as clustering.

Firstly, the students' conversation semantic data are read, and all the data are segmented. The segmentation results are intersected with the emotional dictionary to obtain a new emotional dictionary. For each student's evaluation, find the emotional words in their evaluation to record their positive and negative and record their categories. If there are positive words, the number of positive words is added 1. If there are negative words,

Table 1. Abnormal students' Emotional Dictionary.

exceptional student	productive vocabulary	Negative vocabulary
Students with special physiology	normal, healthy, etc.	Disability, body position information, diabetes, vertigo, gender cognitive ambiguity, etc.
Family abnormal students	Family normal, harmony, happiness, etc.	Single parent, divorce, domestic violence, parents disability, left-behind children, orphans, ethnic minorities, etc.
Economic abnormal students	Rich families, general economic families, etc.	Minimum living security, filing cards, martyrs' children, children of disabled people, subsistence allowance households, poor counties, rural areas, farming, and family per capita annual income are less.
Special academic students	Careful class, complete homework, high test scores, etc.	Subject, credit is not enough, do not understand the class, homework, academic pressure, professional interest.
Psychological abnormal students	Emotional stability, optimism, happiness, personality integrity, etc.	Autism, depression, mania, schizophrenia, anxiety, pain, confusion, loss of love, two-way emotional disorders, suicide, self-mutilation, meaningless life, insomnia, medication, etc.
Special communication students	Confidence, generosity, good manners, rigorous thinking, etc.	social fear, autism, fear of being concerned, please personality, fear of opposite sex and strangers; only I respect, I am me, I am right. etc.

Note: The table above lists only some representative words

the number of negative words is added 1. If there are degree adverbs before the emotional words, the emotional word value is multiplied by different coefficients according to the types of different adverbs. Finally, each student ' s different types of text sentiment values are counted, and the different types of sentiment values are calculated as follows:

$$Emotion_{sentence} = \sum V_{pos} + \sum V_{neg} \tag{1}$$

V_{pos} denotes the weight before positive words, V_{neg} denotes the weight before negative words.

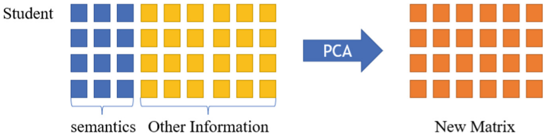

Fig. 1. Dimension reduction of feature vector

3.1.2 Students Data Enhancement and Feature Dimension Reduction

The emotional dictionary method can transform the conversation data of different students into six-dimensional data features. In this paper, the six-dimensional vector is spliced with digital data such as latitude and longitude (family address), psychological test scores, family annual income, number of people in the family, total scholarships and number of contests as the overall feature vector of students.

In order to make the data of different dimensions have the same influence on the results as possible, it is necessary to normalize the data by MIN-MAX, and reduce the dimension of the students ‘ feature vector to 6 dimensions by PCA principal component analysis (Fig. 1).

To facilitate model training, we use 80% of the special student library as our training set, 10% as validation set, and the remaining 10% as test set. In order to solve the problem of large difference in the number of normal samples and abnormal samples, this paper enhances the data of different abnormal data, so that the number of different types of abnormality is roughly equal to that of normal students, so as to improve the accuracy of the model training results.

3.2 Identification of Abnormal Students by Integrated Learning Model

This paper uses the combination of clustering method and neural network to detect abnormal students, and improves the accuracy of the detection of abnormal students by adjusting the parameters of the two models on the verification set.

3.2.1 Student Feature K-Means Clustering

We use K-Means clustering to cluster the feature vectors of students [5, 6]. The number of clusters is set to 6, and the results of each cluster are saved after 10 clustering. The clustering effect is shown in Fig. 2. In order to facilitate the display, Fig. 2 only uses part of the student feature vector as the display.

For the points without labels in the clustering results, we can get the possibility of different abnormal situations. The probability calculation method of the first point belonging to the k-cluster anomaly is as follows:

$$Prob_i^k = \frac{Points^k}{Points} \tag{2}$$

Among them, $Points^k$ denotes the number of points belonging to class k known in the cluster where the point is located, and Points denotes the number of students with known student labels in the cluster where the point is located.

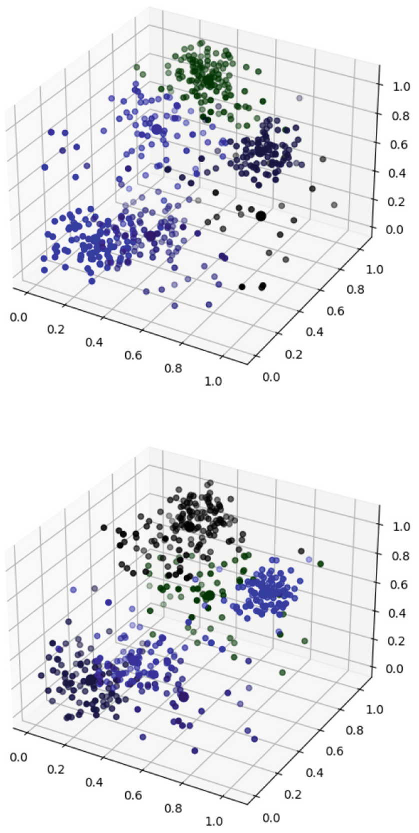

Fig. 2. Student Feature Vector Clustering

The final probability of each student type is the average of 10 clustering results. So far, the possibility of each student belonging to different types of students is obtained by clustering, and the type with the highest possibility is selected as its first prediction label, and its accuracy is calculated on the test set.

3.2.2 Student Category Recognition Based on Neural Network

We use the four-layer fully connected BP neural network as another input model of ensemble learning, in which the eigenvalues of the six dimensions of the students are used as the input, and the possibility of students belonging to different types is used as the output. For the data with real labels, the possibility of the corresponding type of the label is 1, and the possibility of the type of label is 0, where the ReLu function is used as the activation function (Fig. 3).

Due to the small number of labeled data in the original data set, this paper first uses a small number of labeled data to pre-train the neural network. After training, the unlabeled data is input into the neural network to obtain a new data label as a pseudo label of the data. Then the data with false labels and real labels are merged to train the network.

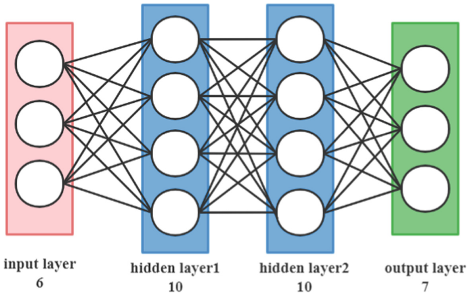

Fig. 3. Neural Network Structure

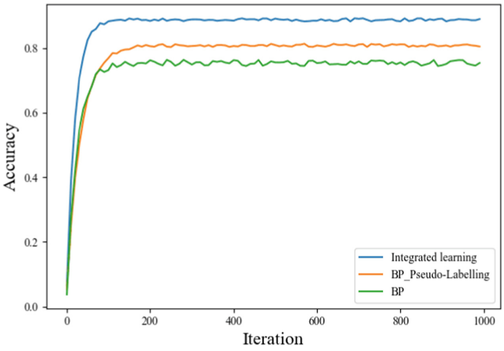

Fig. 4. Accuracy of different models

After the network training is completed, the feature vector of the student is input into the network to obtain the possibility vector of the student belonging to different categories. Similarly, the category with the highest possibility is selected as its second prediction label and its accuracy on the test set is calculated.

3.2.3 Model Integration and Results Display

The possibility matrix P_1 of different categories of students is obtained by clustering method, and the possibility matrix P_2 of different categories of students is obtained by neural network. Assuming that the final possibility matrix of different categories of students is P, the calculation method of P is as follows:

$$P = \lambda_1 * P_1 + \lambda_2 * P_2 \tag{3}$$

By adjusting the values of λ_1 and λ_2, the final P is optimal on the verification set. The accuracy changes of different models in the neural network iteration process are shown in Fig. 4. The graph shows that the accuracy of the special student recognition model based on clustering-neural network ensemble learning in this paper is 82.01%, and the recognition of abnormal students is better realized. With the update of the database, the model can effectively realize the recognition of abnormal students.

4 Conclusions

Colleges and universities bear the mission of educating people for the Party and the country, and abnormal students need more attention. In this paper, a recognition model for abnormal students is proposed. Firstly, abnormal students are calibrated and classified based on working experience. In view of the problems such as huge data of college students, too many text data and difficult to quantify, and difficult to identify abnormal students, the semantic library is used to identify the relevant text information of students, and the emotional feature vector is obtained. Combined with other digital information of students, the feature vector of students is assembled. After the principal component extraction, the abnormal students are accurately identified by combining clustering and neural network. Based on the author' s database, the recognition accuracy of this model can reach 82.02% at present, and the recognition of abnormal students can be realized with data updating.

References

1. Li L.(2021). Analysis of ideological and political education path of college students in post-epidemic era. J. Office Service,16,35+71.
2. Meng Q. (2020). Research on the adjustment of ideological and political education direction of college students in the post-epidemic era. J. 36. 46–51.
3. Wei Z, Jing L.(2021). New characteristics of ideological and political education of college students in post-epidemic era. J. Journal of Liaoning University of Technology. 23. 111-114.
4. Xiaoli W, Mengle J. (2021). Keeping integrity and Innovating -- Exploring the way to run special "Ideological and political Courses" well for special college students. J. Journal of Jingdezhen College. 36. 129-132.
5. Mingyu Z, Jian C, Sutong W, etc. (2021). Analysis of Student educational Portrait based on K-Prototype clustering. Journal of Dalian University of Technology(Social Sciences). 42. 22–31.
6. Ziwei Z, Zihan L, Huixn L, etc. (2021). Portrait visualization of non-cognitive students for learning success. J. Modern Educational Technology. 31. 94-102.