Artificial Intelligence, Medical Engineering and Education
Z.B. Hu et al. (Eds.)

doi:10.3233/ATDE231330

Mental Health State Prediction Method of College Students Based on Integrated Algorithm

Na GUO[1]
College of Resources and Environment, Wuhan University of Technology, Wuhan, Hubei 430070, China
ORCiD ID: Na Guo https://orcid.org/0000-0002-7677-6820

Abstract. Psychological health is an important issue faced by college students, therefore conducting relevant research is meaningful. The use of Adaboost algorithm for ensemble learning, combined with the application of decision tree algorithm, can fully utilize the information in mental health test data and improve the prediction accuracy of the classifier. The C4.5 decision tree algorithm is a commonly used classification algorithm that can classify and distinguish samples based on feature attributes, so it has been selected as the basic algorithm for this study. In order to verify the effectiveness of this method, we selected the mental health test data of 2780 students from a certain university in 2020 for the experiment. Through analyzing experimental results, we found that the method can accurately identify sensitive psychological problems among students. In practical applications, this method can serve as an auxiliary tool to help schools accurately understand the distribution of students' mental health problems, and thus develop corresponding educational measures and intervention plans. In summary, the mental health prediction method based on Adaboost algorithm proposed in the article, combined with the application of decision tree algorithm, can effectively identify psychological problems among college students. In the experiment, this method demonstrated high accuracy and robustness.

Keywords. Adaboost algorithm; University students; Mental health; Prediction methods; Decision trees

1. Introduction

With rapid socio-economic development and increasingly fierce competition in employment, contemporary university students are facing unprecedented pressure from academic, employment and interpersonal relationships, which has led to the emergence of psychological disorders and the occurrence of malicious incidents. Students are more prone to mental health problems such as confusion, anxiety, anxiety, low self-esteem, rebellion and even hate psychology due to their studies and employment. Therefore, the use of technology to identify the mental health problems timely and to take measures to educate, guide and treat them has become an important topic of research in mental health education in higher education. Data mining is a knowledge discovery process that uses algorithms to search for hidden information from a large amount of

[1] Corresponding Author: Na GUO, E-mail: 234074489@qq.com.

data, allowing intelligent and efficient analysis and processing of big data. The application of data mining methods to the analysis of university students' psychological problems and crisis warning has attracted widespread attention from researchers in China.

2. Related Technical Studies

The level of individual mental health has always been a major variable in shaping and promoting their growth and development [1]. China has clearly stated the importance of psychological counseling and improving mental health education. In recent years, relevant policies on improving public mental health have been continuously improved and strengthened to enhance public awareness of mental health. However, the problem of public mental health has not been fully resolved [2]. With the further attention of the state and people on this issue, more and more experts in the field of psychology have invested in the research of mental health problems [3]. They gain a deeper understanding in the field of the treatment and prevention of mental health problems, and in the continuous research, the methods and models of mental health prediction have been also continuously developed. However, the research on mental health still cannot keep up with the pace of social development, and there is a problem of asynchronous development [4]. So far, domestic experts' research on mental health mainly in facing psychological problems, coping strategies, challenges, reasons and how to predict and deal with them in time. Some scholars mainly analyze the root causes, solutions, and challenges faced by current college students' mental health problems [5]. For the prediction of psychological problems, two main methods are used, one is based on research on network user data.

They analyze user behavior to predict user psychological states by obtaining network behavior data of APP users, such as Weibo and Twitter [6]. For example, through analyzing students' network behavior data and using support vector machine algorithm, random forest algorithm, etc., they can predict students' psychological health. Using Weibo active user data, based on the Big Five personality assessment scale, the personality variables can be classified and predicted through robust multi-task learning model. Through data analysis from Weibo data, they can predict risks using MLP model. However, there is a problem in these studies, that is, only through users' network behavior for prediction, so these prediction results may not be accurate [7].

Another method is to directly obtain first-hand psychological data through designing and distributing mental health questionnaires, and then analyze these data to predict mental health. Questionnaire surveys are more direct and accurate than the first method, without further analysis of the underlying psychological state of the user. For example, using multiple regression analysis to analyze the impact of positive psychological traits on mental health in college students; conducting questionnaire surveys among freshmen, and using a BP neural network model to predict students' mental health [8].

The core idea is to adjust the distribution probability of each sample in a set of psychological data to form different training sets, and integrate these basic classifiers using different weights to generate a powerful classifier. As the number of iterations increases, the upper limit of the training error rate decreases, while effectively avoiding the overfitting problems common in other algorithms, thus improving the classification accuracy of the classifier [9].

3. Mental Health Prediction Model

The decision tree algorithm is a tree-structured classification and regression method that combines classification and regression. As a classification method, it is based on the classification of feature attributes and builds a model based on the principle of minimizing the loss function. until it reaches the leaf node, and the category in the corresponding leaf node is the decision result of the model. In a decision tree, only the leaf nodes can represent the final classification, with different leaf nodes representing different classifications and other non-leaf nodes representing features or attributes [10]. The ultimate goal is to be able to correctly classify the dataset. As the model structure is a tree structure, the classification decision is made through layers of conditions, which is intuitive and easy to understand, and has the advantages of high accuracy and simple model. The decision tree algorithms used commonly include ID3 and C4.5, etc. In this paper we will use the C4.5 algorithms. The classification weights are used as parameters to adjust the sample weight distribution D. After several iterations of training, the classification of each basic classifier and its weights are obtained. The model is shown in Figure 1. The mental health prediction model is shown in Figure 1.

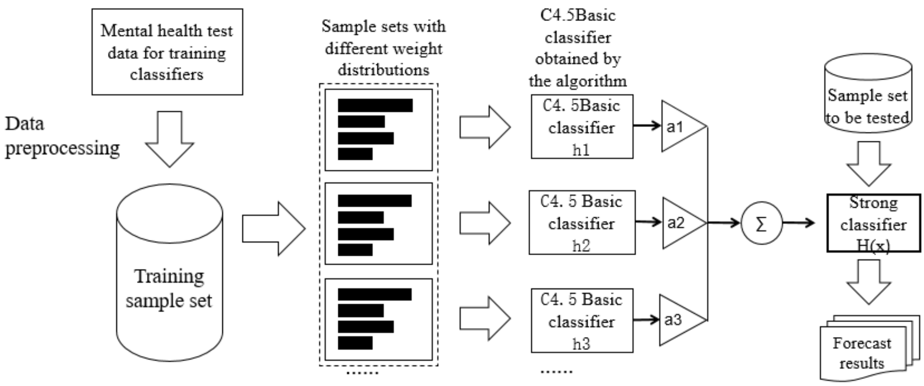

Figure 1. Mental health prediction model

4. Key Technologies

4.1. Basic Classifier Generation Algorithm

In this article, we used the C4.5 decision tree algorithm, which is a basic classifier generation algorithm. The decision tree model is a rule-based reasoning model that can learn from training samples, establish classification rules, and then classify new samples. The information gain rate is used in the C4.5 decision tree algorithm as the criterion for selecting branch attributes, which solves the problem of simply using information gain that tends to prefer attributes with more values. In addition, C4.5 also has the ability to handle incomplete and non-discrete data. Incomplete data refers to the presence of missing values in the training set, while non-discrete data refers to attributes that are not only discrete values, but also include continuous values or other forms of data. In summary, the C4.5 algorithms, which is regarded as an improved decision tree algorithm, uses information entropy and information gain rate for classification, and has high efficiency and flexibility in processing non-numeric data, selecting branch attributes, and handling incomplete and non-discrete data. By the decision tree model generated in the C4.5 algorithms, we can better classify new samples.

The following is a new description of the specific steps of the C4.5 decision tree algorithm:

(1) Initialize the starting node of the decision tree, also known as the root;

(2) If all samples S belong to a certain category C, return Root as the leaf node and label it with category C;

(3) For each element A∈ in the tribute Set, calculate its information gain ratio (A);

(4) Select the attribute Set with the largest Gain ratio(A) value as the test attribute for Root, that is, the optimal segmentation attribute Attest;

(5) This new leaf node will be marked as the category with the largest number of samples included in the node.

4.2. Adaboost Process for Constructing Strong Classifiers

The Adaboost algorithm is an improved version of the Boosting algorithm, aiming to obtain a high-precision strong classifier by integrating multiple weak classifiers (also known as basic classifiers). In the process of constructing a strong classifier, this article adopts the Adaboost algorithm and follows the following steps. First, we need to perform preprocessing to obtain training samples for mental health testing data. This can include data cleaning, feature selection, and feature transformation steps to ensure that the training data used has a certain quality and accuracy. Next, we repeatedly call the C4.5 algorithms to generate a series of basic classifiers. The C4.5 algorithms is a decision tree algorithm that can generate a decision tree model based on given training data. Each time we call the C4.5 algorithms, we get a different basic classifier that performs classification under different feature combinations and decision rules. Then, for each basic classifier, we assign a weight based on its correct classification rate. Classifiers with high correct classification rates will be given higher weights, while those with low correct classification rates will be given lower weights. The purpose of this is to make those classifiers that perform well in the training data play a greater role in the final strong classifier. Finally, we combine all the basic classifiers to obtain the final strong classifier. When some new samples are detected, we provide these samples in parallel to all the basic classifiers, and add the weights of those classifiers that have the same classification result. Finally, the highest weight of the classification result will be chosen as the output of the strong classifier. This method of using Adaboost to build a strong classifier can improve the accuracy and stability of classification. By integrating the prediction results of multiple basic classifiers, we can leverage their respective strengths to obtain a more reliable and accurate classification result.

The logistic regression model is solved using a gradient ascent/descent algorithm for optimization. If you want to minimize the loss function you use gradient descent, if you want to maximize the likelihood function you use gradient ascent, they are essentially the same thing. Firstly, a sample set containing N training samples of mental health data is inputted.

$$D=\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_N,y_N)\}\ x_i\in X,y_i\in Y \tag{1}$$

X is the sample set, Y is the sample class, $Y = \{-1, +1\}$ and each sample x_i contains the K-dimensional feature .

In the second step, the weight distribution of the sample set is initialized.

$$D_1 = \{w_1(x_i)\}, w_1(x_i) = \frac{1}{N}, i = 1, 2, \cdots, N \tag{2}$$

In the third step, perform a loop on any $t \in \{1, 2, \cdots, T\}$ (where T is the number of basic classifiers).

Using samples with a weight distribution of D_t for learning, a basic classifier h_j is trained for each $v_j(x)$.

$$h_j = \begin{cases} 1, p_j v_j(x) < p_j \theta_j \\ -1, p_j v_j(x) .. p_j \theta_j \end{cases} \tag{3}$$

θ_j is the threshold value, p_j is the bias value and $p_j \in \{-1, 1\}$.

The weighted error rate is calculated.

$$\varepsilon_j = \sum_{i=1}^{n} w_t(x_i) \left| h_j(x_i) \neq y_i \right| \tag{4}$$

The error rate ε_j minimum value h_j corresponding to ε_t is used as the basic classifier for this loop h_t

(2) Calculation of the weighting parameter α_t for h_t.

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right) \tag{5}$$

(3) Update the weighting distribution using the weighting parameter α_t.

$$D_{t+1} = \{w_{t+1}(x_i)\}, w_{t+1}(x_i) = \frac{w_t(x_i) e^{(-\alpha_t y_i h_t(x_i))}}{Z_t} \quad i = 1, 2, \cdots, N \tag{6}$$

Z_t For the normalization factor, $Z_t = \sum w_t(x_i) e^{(-\alpha, y, h_t(x_i))}$.

In the fourth step, the final strong classifier is constructed through the T optimal basic classifiers obtained from the T rounds of training $H(x)$.

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right) \tag{7}$$

5. Experiments and Analysis of Results

This study collected data from students in the class of 2015 in a university to test their mental health and personality traits. The mental health test used the SCL-90 scale, including 90 items divided into 10 factors. Each factor has 5 items, and participants are required to choose 1 (none), 2 (very slight), 3 (moderate), 4 (serious), or 5 (extremely serious) to indicate the degree of psychological symptoms of the corresponding item based on their own situation. In addition, the personality test used the Eysenck Personality Questionnaire (EPQ). The questionnaire consists of four scales. Through the collection and analysis of these tests, we will be able to gain a deeper understanding of the mental health status and personality traits of the college student population. This will help provide them with better mental health support and intervention strategies to promote their all-round development. The first three scales represent three separate dimensions of personality structure, while L is the validity scale, which represents fictive personality traits and the level of social simplicity and infantilism. The version of the scale used in this paper is the Eysenck Personality Questionnaire Short Form Chinese version (EPQ- RSC).

Model evaluation plays an important role in the entire data mining process [11-13]. Before each prediction model is implemented, it is difficult to know the specific strengths and weaknesses of the prediction model in the research direction. Therefore, multiple prediction model experiments are required, and model evaluation metrics are used to evaluate each model based on the prediction results, in order to select the optimal prediction model for the research direction. As the possible response options for a scale entry are discrete values, missing values are filled in by counting the most frequent values for that entry. To facilitate data processing by the algorithm, the values taken for all questionnaire items are abstracted into codes.

The experiments were carried out on a Dell Inspiron 3650-D1838 machine and the algorithms were programmed in Python. The test samples were divided into a training set and a test set. Of these, 1,800 were training samples and 980 were test samples. The Adaboost integrated learning classifier was trained using the training samples, and the positive and negative samples were classified based on the expert assessment of the samples by the school's mental health counselling center. The final test set was tested and compared with the Mental Health Assessment Centre's assessment data. To observe the final performance of the classifier, the following two indicators were considered.

$$\text{Recall} = \frac{PP}{PP + NP} \tag{8}$$

The final classifier was tested on 980 test samples and the results were compared with the results of the basic classifier based on the decision tree algorithm C4.5 on the same dataset and the statistics are shown in Table 1.

Table 1. Classifier classification results

Classified ware	Number of positive samples	Negative sample size	PP	PN	NP	NN	Recall	Accuracy
Adaboost + C4.5	28	952	26	12	2	940	0.9286	0.9857
C4.5	28	952	25	27	3	925	0.8928	0.9694

The results in Table 1 indicate that the performance of the classifier based on the integrated learning Adaboost algorithm has improved significantly in terms of accuracy and recall [14].

6. Conclusion

This article proposes a method that combines the C4.5 decision tree algorithm and Adaboost ensemble learning algorithm to construct a classifier for predicting the mental health problems that college students may face. The experimental results show that the ensemble learning classifier using this algorithm has significant improvements in both accuracy and recall. This algorithm can provide support to mental health counselors, student administrators, and counselors understand students' psychological development, main symptoms of mental health problems, and personality traits, enabling them to focus on students who may have mental health problems, provide deeper care, guidance, and psychological therapy, and prevent them from encountering mental health problems in the future.

The research of the college students' mental health is an important and promising field. With the increasing complexity of the social environment, college students are facing increasing psychological stress [15]. With the ubiquity of mental health issues, many researches are paying more and more attention to the development of mental health education curriculum, and exploring how to integrate it into the university curriculum. The rapid development of digitization and information technology has made online psychological counseling possible, and this form has been applied in many places. This article hopes to provide services to more college students who need help through online platforms. Depending on different sources of stress, college students' coping strategies will differ; this article hopes to find the most effective coping strategies to help college students deal with various pressures. For those special groups that are easily affected by psychological stress, this article hopes to gain a deep understanding of their mental health status in order to provide effective intervention and help. In general, research in the field of college students' mental health is constantly expanding to meet society's needs. In the future, we may see more innovative research methods and strategies being used to improve the mental health of college students.

Reference

[1] Prodipto Bishnu Angon, Sujit Mondal, Chandona Rani Das, Mintu Kumar Bishnu. Behavioral Changes of Children Intelligence for the Extreme Affection of Parents [J]. International Journal of Education and Management Engineering (IJEME), 2021, 11(6): 20-28.

[2] Yu Jiayuan. Application of fuzzy comprehensive evaluation based on genetic algorithm in psychometric measurement [J]. Journal of Psychology, 2009(10): 1015-1023 (in Chinese).

[3] Ahuja, R., & Banga, A. Mental stress detection in university students using machine learning algorithms [J]. Procedia Computer Science, 2019, 152: 349–353.

[4] Kim, J. Y., Liu, N., Tan, H. X., & Chu, C. H. Unobtrusive Monitoring to Detect Depression for Elderly with Chronic Illnesses [J]. IEEE Sensors Journal, 2017, 17(17): 5694–5704.

[5] Andrews, G., Bell, C., Boyce, P., Gale, C., Lampe, L., Marwat, O., Wilkins, G. Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for the treatment of panic disorder, social anxiety disorder and generalised anxiety disorder [J]. Australian and New Zealand Journal of Psychiatry, 2018, 52(12): 1109–1172.

[6] He Guangdong. Research on the application of data mining technology in the analysis of college students' psychological problems [D]. Baoding: Hebei University, 2013 (in Chinese).

[7] Muruka C. and Muruka, A. Guidelines for Environmental Health Management in Children's Homes in Sub-Sahara Africa [J]. International Journal of Environmental Research and Public Health. 2007, 4(4): 319-331.

[8] Alan Abelsohn, John Frank and John Eyles. Environmental Public Health Tracking/Surveillance in Canada: A Commentary [J]. Healthcare Policy, 2009, 4(3): 37-52.

[9] Adebayo P. Idowu, Emmanuel R. Adagunodo, Olapeju A. Esimai. Development of a Web Based Environmental Health Tracking System for Nigeria [J]. International Journal of Information Technology and Computer Science, 2012, 4(7): 61-71.

[10] Nor Safika Mohd Shafiee, Sofianita Mutalib. Prediction of Mental Health Problems among Higher Education Student Using Machine Learning [J]. International Journal of Education and Management Engineering, 2020, 10(6): 1-9.

[11] Snigdho Dip Howlader, Tushar Biswas, Aishwarjyo Roy, Golam Mortuja, Dip Nandi, "A Comparative Analysis of Algorithms for Heart Disease Prediction Using Data Mining", International Journal of Information Technology and Computer Science, Vol.15, No.5, pp.45-54, 2023.

[12] Md. Al Muzahid Nayim, Fahmidul Alam, Md. Rasel, Ragib Shahriar, Dip Nandi, "Comparative Analysis of Data Mining Techniques to Predict Cardiovascular Disease", International Journal of Information Technology and Computer Science, Vol.14, No.6, pp.23-32, 2022.

[13] P. Amaranatha Reddy, MHM Krishna Prasad, " Sliding Window Based High Utility Item-Sets Mining over Data Stream Using Extended Global Utility Item-Sets Tree", International Journal of Image, Graphics and Signal Processing, Vol.14, No.5, pp. 72-83, 2022.

[14] Sabourin, A. A., Prater, J. C., & Mason, N. A. Assessment of mental health in Doctor of Pharmacy Students [J]. Currents in Pharmacy Teaching and Learning, 2019, 11(3): 243–250.

[15] McLafferty, M., Lapsley, C. R., Ennis, E., Armour, C., Murphy, S., Bunting, B. P., O'Neill, S. M. Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland [J]. PLoS ONE, 2017, 12(12): 1–14.