Automatic Assessment Method of College Students Psychological Stress Based on Medical Big Data

Xiang Li(✉)

Jingchu University of Technology Normal University, Jingmen 448000, Hubei, China
qihang7895@126.com

Abstract. Students' psychological stress assessment has become an important part of college mental health management, but the current assessment methods are qualitative analysis, combined with the scale to get the assessment results. The lack of data support in the evaluation process leads to large evaluation errors. In view of the above practical problems, this paper studies the automatic evaluation method of college students' psychological stress based on medical big data. After using crawlers to obtain medical big data, it was preprocessed. Taking physiological data as the evaluation index of psychological stress, the relationship between physiological indexes and psychological stress is obtained by data mining. The HMM model improved by SVM is used to quantify the evaluation results. After testing, the relative error of the evaluation method based on medical big data is less than 10%, and the evaluation accuracy is higher.

Keywords: Medical Big Data · College Students' Psychology · Psychological Stress · Evaluation Methods

1 Introduction

The acceleration of globalization, the rapid development of information and the openness of information increase the psychological pressure of college students. Academic, social relations, employment and other kinds of pressure on college students exhausted. When people encounter stressful events in their lives, there is a sense of stress. Psychological stress is a kind of psychological stress reaction when people face difficult situations [1]. College students also face more stressful events in complex environments. Stress and other psychological factors can affect physical health, leading to poor occupational health and insomnia, anxiety, mild depression and other physical and mental disorders. Studies have shown that appropriate levels of pressure can motivate people to progress and reach their potential. For students, proper pressure can improve their learning efficiency and benefit their growth and development. However, excessive psychological pressure may bring students physical and psychological pain. When people do not know how to deal with this kind of pressure, there will be a variety of negative emotions. Students can not solve the psychological pressure when prone to psychological problems,

S. Wang (Ed.): IoTCare 2022, LNICST 501, pp. 61–72, 2023.
https://doi.org/10.1007/978-3-031-33545-7_5

leading to accidents. Psychological stress is a more obscure topic, many students may not even notice their own situation has been very serious or deliberately ignored it. Even if students are in a state of abnormal psychological stress, because of the convergence of character rarely find others to express, which caused the school to monitor students' psychological stress. College students' mental health has become the most concerned problem in colleges and universities. Effective ways are needed to regulate students' psychological pressure and negative emotions.

Students with different levels of psychological stress can choose different ways of adjustment. Therefore, evaluating students' psychological stress is the basis of dealing with college students' psychological problems. In recent years, there are many researches on mental health of college students, but few researches on mental stress assessment of college students. The traditional psychological stress assessment scheme is mainly realized by questionnaire and manual interaction. Psychological stress was assessed by daily or multiple questionnaires and manual interviews over a period of time, and by questionnaires and wearable equipment. This approach usually requires a relatively long period of time for data collection and works only for some of the students involved in the survey [2]. Using the recorded information to predict, do not need to re-operate each survey, and after the model is built, can be extended to all students with similar records, adaptable. But choosing the right features and the right models will greatly affect the assessment [3].

With the progress of the times, the problem of psychological stress of college students has attracted people's attention more and more obviously, and it has entered the public's field of vision. It belongs to the important and practical field of the branch of educational data mining. Traditional psychological stress assessment programs are mainly realized through questionnaires and human interaction. Some use daily or multiple questionnaires and manual interviews for a period of time to assess students' psychological pressure, and some use questionnaires and wearable devices. To assess the psychological stress of students, this method usually requires a relatively long period of time to collect data, and it is only effective for some students who participated in the survey. Under the background of big data era, with the development of health care and intelligent medical care, medical information is growing exponentially every day. These accumulated medical big data contain rich information and inestimable value. With the development of key technologies such as distributed storage and distributed computing, it is easy to solve the difficult problems in traditional data warehouse. Compared with questionnaire survey data, medical big data is more objective and accurate, because it records students' real medical feedback information and diagnosis and treatment behavior. According to the above analysis, this paper will study the automatic evaluation method of college students' psychological stress based on medical big data. Use crawler technology to crawl medical data related to psychology. The medical data is then preprocessed. Through data mining technology, the relationship between physiological data and psychological stress scale was established. On this basis, automatic evaluation of psychological stress of college students is carried out. Through the big data analysis technology, the paper quantitatively evaluates the psychological stress of college students, and provides empirical data support and reference.

2 Medical Big Data Acquisition and Analysis Processing

2.1 Crawlers Acquire Medical Big Data

The rapid development of medical information industry has given birth to a considerable scale of medical platform. These medical platforms are dedicated to provide convenient medical services for ordinary users, while facilitating access to medical resources, but also accumulated a large. The medical data of college students in psychological counseling or other medical behaviors are mixed with the medical data of different groups, which affects the efficiency of follow-up analysis. Therefore, crawlers were used to crawl medical data related to college students' psychological stress assessment.

Map/Reduce is used to construct intelligent web crawler for medical big data. The crawler for medical big data includes crawler strategy design, middleware design and data storage design. The design of crawler strategy mainly involves how to parse and extract web page information, incremental crawling data and data deduplication, etc. (mainly using Redis database to select efficient mode to assign captured URL through custom module. According to the priority of crawling data and ensuring the consistency of crawling order, the crawling URL order strategy is configured. The part of middleware design mainly deals with the anti-crawler setting based on Scrapy crawler framework and the design of appropriate proxy pool for IP real-time replacement [4].

Due to the wide availability of medical big data. This study uses the distributed master-slave crawler architecture shown in Fig. 1 below to obtain medical big data of college students.

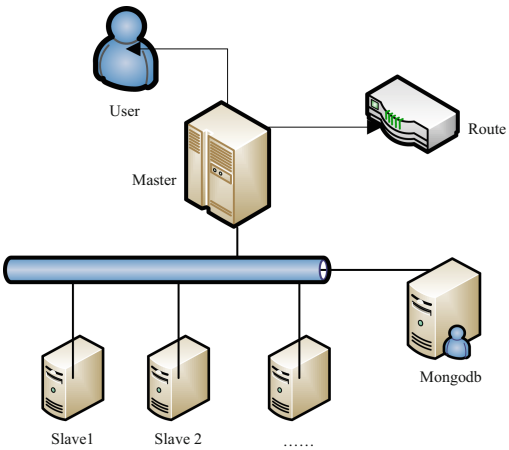

Fig. 1. Schematic diagram of distributed crawler framework

As a host, Master is responsible for scheduling crawler tasks and managing cluster resources. It also assigns crawler requests to RPC communication between Slave processing and master-slave machines for real-time monitoring of crawler tasks. Each of these Slave receives a crawler task assigned by the Master and processes the task only on its own node, storing the results in the MongoDB database. The Redis database stores

the status of the crawl URL queue, which is fetched from the cache by the host and slave. The second stage mainly includes deciding whether to crawl the page. If the task is not crawled directly into Hbase for parsing, crawl the content of the page for content parsing. At the same time, the output of the parsed content is saved, and it is optional whether the parsed results should generate the next level of tasks [5]. The third stage is to store the output data, and the source data crawled after parsing the content will be stored in the mongodb database, Hbase database, and test according to different requirements. Where test is custom and optional to test the accuracy of the data.

The complete process for crawlers to obtain data on college students' psychological stress is shown in Fig. 2.

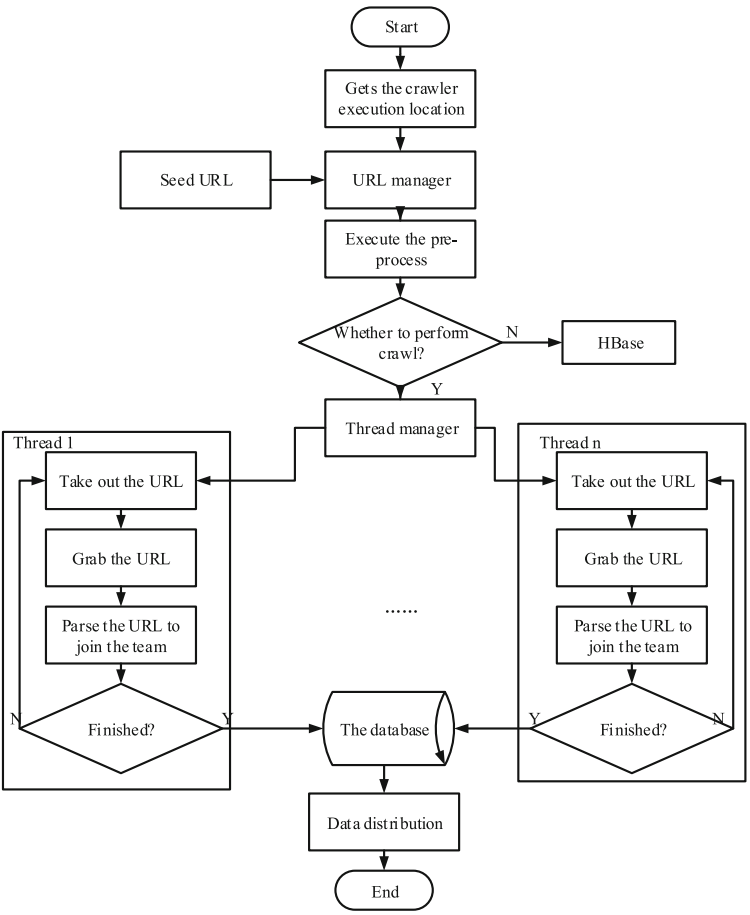

Fig. 2. Obtaining the crawler process of college students' psychological stress medical big data

Task start execution to get crawler execution configured as the first phase. This phase is the crawler task generation phase. This phase implies a timed task module that pre-generates crawler tasks based on the crawler configuration. In order to avoid the web

crawler from falling into the web trap, the authoritative web pages related to medical subjects are selected as part of the seed set.

This design uses Bayesian algorithm to identify crawler crawling objects Web pages. To identify whether the text on the Web page contains medical data related to psychological stress assessment of students. Calculate Bayesian probabilities for uncategorized texts based on setting up texts for medical big data related words [6].

$$P(L_i|w) = \frac{P(w|L_i) * P(L_i)}{\sum_{i=1}^{n} P(w|L_i) * P(L_i)} \tag{1}$$

In the above expression, $P(L_i|w)$ indicates the probability that the document belongs to category L_i, determined by the matching of its words with vector pattern w. Posterior probability $P(L_i|w)$ is obtained by prior probability $P(L_i)$ and conditional probability $P(w|L_i)$. By finding the maximum value of $P(L_i|w)$, you can know the classification of web pages, namely:

$$F(w) = \arg\min_{L_i \in L} P(L_i|w) \tag{2}$$

It is assumed that the medical big data texts used for classification are independent of each other. Follow the process above to categorize the URLs. According to the maximum classification results, after identifying the target web, the crawler crawls the corresponding web page. Web crawlers crawl files in a variety of formats, including Word, PDF, Excel, and so on, and the corresponding Jar packages in Java provide interfaces for processing these files. Web crawlers need to call the corresponding file parsing module based on the file extension to parse these binaries. Then the index program is used to index the analyzed web page information, and the index database is established. Finally, for user search.

2.2 Medical Big Data Preprocessing

When preparing data, be fully aware that real data is susceptible to interference, loss, and inconsistency. So the first thing to do when you get a data source is to preprocess the data to make it a high-quality data mining source. The primary function of data preprocessing is to transform raw data into a data format that can be entered into the model. It includes data cleaning, data integration, data specification and data transformation.

In medical big data processing, we often meet the problem of imbalance of positive and negative labels. If we do not deal with it, we will get a defective evaluation model. The main algorithms for dealing with imbalanced datasets are as follows: (1) Adjust the threshold value to make more samples become samples with fewer categories. (2) Cost-sensitive learning makes the category training with fewer samples get higher weight, which makes the punishment for wrong judgment greater. (3) Oversampling, copying more small categories of samples and equalizing the overall sample ratio. (4) Undersampling reduces the number of samples with a large number of categories, resulting in a balanced ratio of the overall sample. (4) Data synthesis (SMOTE), i.e., synthetic minority over-sampling technology. The algorithm will feature a small number of samples and artificially synthesize new samples to add to the dataset [7].

The value interval of data is normalized, and the projection space [0, 1] is generally taken as the projection space. Therefore, the weights of different ordinal variables are unified so as to facilitate the calculation of similarity and clustering operation. For the acquired software running dataset, the normalized formula for its data mapping is as follows:

$$S_{R-D} = \frac{R_D - 1}{\max R - 1} \tag{3}$$

In the above expression, R_D is the data whose ordinal variable attribute is k. $\max R$ is the maximum value for the data map. In order to adapt to the range of the degree of variability of different types of variables, it is necessary to transform the degree of variability of different types of data. Make their values map to the same interval [0, 1].

The LOF algorithm is used to detect outliers in the data to reduce the impact on subsequent analysis. The neighborhood point set $l_k(p)$ of data point p is the accessible distance from all points to p. Rather than the accessible distance of all points p to $l_k(p)$. The lower the data point density in the neighborhood of data point p, the more likely it is an outlier. Conversely, the higher the density, the more likely it is that the points belong to the same cluster. If the data point p and the surrounding neighbor point density is more sparse. Then the accessible distance of the point has a large probability to take a larger value, resulting in a smaller data point density in the neighborhood of data point p. If the data points p and surrounding neighborhood points are more dense. Then the accessible distance may be a smaller distance, and the density of data points in the set of data points p is higher [8].

$$LOF(p, o)_k = \frac{\sum_{q=1}^{N} q l_k(p) l_k(o)}{|l_k(p)|} \tag{4}$$

In the above formula, o is the data point in the neighbor point set $l_k(p)$ of the data point p. $l_k(o)$ is the maximum distance from the data point p in the domain set. If the value calculated by the above formula is closer to 1. Then it means that the density of point p and its neighbor set $l_k(p)$ is similar, and the more likely this point is a normal point.

3 Analysis on the Evaluation Index of College Students' Psychological Stress

College students will encounter various pressures in their daily life. When these pressures accumulate to a certain extent, it will affect students' life satisfaction level. These external pressures are objective factors, as well as subjective factors that affect an individual's attitude towards and satisfaction with life. External stressors have a direct effect on individual mental health and life satisfaction. When the individual psychological stress is too high, it will lead to changes in their physiological data. Therefore, this paper takes the physiological indexes in the big data of medical treatment as the evaluation basis, and gets more accurate evaluation results by quantification.

EEG signals are closely related to physiological and psychological information related to various parts of the body. In general, the excitation process of EEG signals increases significantly when the slow wave with high amplitude changes to the fast wave with low amplitude. Conversely, suppression occurs when a low amplitude fast wave becomes a high amplitude slow wave. When neurons in the brain are active, they exhibit nonlinear dynamics. Therefore, nonlinear dynamic analysis methods such as complexity, psychological stress index and maximum Lyapunov index (LLE) are combined. They can accurately analyze the nonlinear dynamic characteristics of EEG signals.

KC complexity is a coarse grained nonlinear dynamic method. It can judge the random degree of the sequence. When there are enough sampled data, the mean of KC complexity tends to 0 in periodic sequence and 1 in random sequence. In other sequences between 0 and 1 [9].

Let EEG sequence be $R = \{r_1, r_2, \ldots, r_n\}$. The new sequences are $R' = \{r'_1, r'_2, \ldots, r'_n\}$ and $R'' = \{r'_{m+1}, r'_{m+2}, \ldots, r'_{m+i}\}$. The KC complexity calculation steps are as follows.

(a) The sequence R' is further constructed from the sequence R, and a value greater than the average value of sequence R is replaced by 1, otherwise 0. To get the new sequence R', and then add a string of characters after the sequence R''.
(b) Compute subsequence $R'R''$ according to the sequences R' and R'' obtained from step (a). If sequence $R'' \in R'R''$, copy the sequence after R' and repeat the calculation. Otherwise, insert a "*" after $R'R''$, and repeat the process until the end of the sequence to get a new sequence.

The formula for obtaining the KC complexity from the sequence constructed by steps (a) and (b) is as follows:

$$\upsilon = \frac{R''(n)}{R'R''(n)} \tag{5}$$

The energy of EEG signals $\delta, \theta, \alpha, \beta$ and γ varies with the fluctuation of brain states. Therefore, this paper combines the characteristics of psychological stress in human physiological data. Based on Hilbert's marginal spectral energy, psychological stress index is defined. The calculation expression is as follows:

$$P = \frac{E\delta + E\theta}{E\alpha + E\beta} \tag{6}$$

Among them, E is the marginal spectral energy corresponding to four kinds of rhythms.

When the human body receives certain outside stimulation, will have the psychological tension or the nerve excited phenomenon, causes human body's heart to relax and the contraction speed to speed up, the heart rate will also increase accordingly. The spectral peak recognition curve calculated by the autoregressive model of heart rate variability is accurate and has high power spectral resolution. By reading the data on the spectral estimation curve, the corresponding interval of heart rate can be determined.

Respiration rate RR is calculated by means of the mean value between the peak periods of the respiration wave. Blood oxygen saturation is one of the important indexes

to measure the oxygen content in human blood. Blood oxygen saturation is the amount of oxygenated hemoglobin in the blood as a percentage of the total binding hemoglobin volume. The above physiological indexes are used as the specific quantitative criteria to evaluate students' psychological stress. The relationship between physiological data and psychological stress scale was established by data mining.

4 Medical Big Data Mining

Medical big data has the characteristics of high-dimensional data. In order to simplify the processing process, a global algorithm model is used to reduce the dimensionality of the data. Given a dataset X consisting of n samples and m features. It is known that it contains C modes, and the label information B of X is known. The data is dimensionally reduced according to the principle of mutual information.

$$D = \arg\max_{d} \left\{ \frac{1}{|D|} \sum_{x_i \in D}^{n} I(x_i; b) - \frac{1}{|D|^2} \sum_{x_i, x_j \in D}^{n} I\left(x_i; x_j\right) \right\} \tag{7}$$

In the above formula, D is the feature selection target. Mutual information is a symmetric metric, that is, the information quantity of feature B obtained by observing feature X is equal to the information quantity of feature X obtained by observing variable $I(X; B) \equiv I(B; X)$.

This symmetry is a good attribute in feature selection. But mutual information tends to have more values. When X or B has more values, the results calculated by mutual information show that the probability of their correlation is very high. So if you compute the mutual information of the features corresponding to many tags, you can easily get a very large value [10].

Symmetric uncertainties are used to compensate for the bias of mutual information to features with more values. In order to measure the identification of feature X to tag B, the correlation evaluation matrix $V = \{DU(x_i, b_i)\}$ is constructed by using the symmetric uncertainty of feature and tag. The value of $V = \{DU(x_i, b_i)\}$ ranges between [0, l].

The larger the value is, the more important the relevance between the representation features and the classification tasks is. Feature redundancy evaluation matrix V^* is constructed by using symmetric uncertainty between features. Combining the above evaluation matrices V and V^* to maximize the feature correlation. At the same time, the feature redundancy is minimized and the objective function is constructed as follows.

$$f = \min_{W} \left\{ W^T V^* W - W^T V \right\} \tag{8}$$

The process of optimizing the above objective function is mapping medical big data from high dimensional space to low dimensional space. After dimensionality reduction, k-means clustering algorithm is used to mine the relationship between data and psychological stress.

Set the densest point to the first initial cluster center point Go through it, calculate the distance from each point d_i of $C1$ in the high density region, calculate the distance between them. Set the point farthest from $C1$ as the second initial cluster center point

$C2$. Taking these two points as the center, two clusters are generated according to the distance between the center and other points, and the cluster centroid c_1, c_2. Find the furthest point from the two clustering centers in the high density domain. This paper sets it as the third initial clustering center point. The formula is as follows:

$$J_i = j(d_i, c_1) + j(d_i, c_2) \tag{9}$$

Among them, c_1 and c_2 represent respectively the center of mass with $C1$ and $C2$. The distance c_1 and c_2 farthest point is chosen as the third cluster center point $C3$. $j(d_i, c_i)$ is the Euclidean distance from the cluster center. Repeat the process until the number of initial cluster centers is equal to the number of predefined clusters to get the corresponding initial cluster centers. According to the process of k-means clustering algorithm, medical big data are clustered.

The following intra-cluster variation is used to evaluate the quality of clustering results.

$$e = \sum_{i=1}^{k} \sum dist\left(d_j, ce_i\right)^2 \tag{10}$$

In the above expression, e is the mean distance difference within the cluster. *dist* is the square of Euclidean distance between clustering centers. ce_i is the cluster center. d_j is medical big data to be mined. Through cluster analysis, the data relationship between medical big data and college students' psychological stress was established. In this paper, the relational features are used as the classification basis, and the HMM model is used to analyze the data to obtain the results of psychological stress assessment.

5 Realize the Psychological Stress Assessment of College Students

This paper uses SVM algorithm to improve HMM model. In the model of college students' psychological stress assessment, SVM chooses "one-to-many" strategy, and HMM chooses continuous strategy. The input parameters in the model are dominant, which represent the probabilistic vector sequence of physiological parameters after SVM, and the output pressure emotion sequence is hidden. On this basis, the stress emotion is classified to improve the accuracy of stress emotion classification.

The training process of HMM emotion model based on SVM to optimize feature parameters is as follows:

(1) SVM training: Take the optimized feature vector sequence as the input vector of this step, rain three SVM, write down three labels, describe as A, B, C. Respectively extract A corresponding vector is positive, B, C corresponding vector is negative. B is positive and A and C is negative. C is positive, B is negative. All the samples belonging to this category are input into three SVM and the output probabilities are calculated respectively to get three probabilities.

(2) HMM training: This step involves training three HMM models $HMM = (\pi, \mathrm{A}, B)$. Each class corresponds to a model, and the training data is three output probabilities in the previous step, that is, the output of each class is a vector (g_1, g_2, g_3). This

is done for all training samples and the vector sequence is obtained. The model parameters are then randomly initialized. BW algorithm is used to calculate the parameters of the model, and stable π, A and B are obtained, which are parametric models related to college students' psychological stress.

HMM emotion model improved by SVM is used to identify the optimized feature parameters. The corresponding parameters of SVM and HMM model are obtained after training stage. First, we input SVM to get the sequence of probabilities, then input three HMMs. The observation probability matrix of the model is estimated by using the piecewise K-means principle. After selecting the best observation probability matrix, Baum-Welch is used to reevaluate the parameters in the model and judge whether the model converges. If the model parameter is convergent, the model under the parameter is used for flutter identification, otherwise, the process is repeated until the best parameter is determined. After HMM calculation, the results of psychological stress assessment were obtained. So far, we have completed the research on the automatic assessment of college students' psychological stress based on medical big data.

6 Test Study

Above proposed based on the medical big data university student psychology pressure automatic assessment method. In order to judge the practical application value of this method, the performance of this method will be tested with data set.

6.1 Test Content

Under the same condition, the evaluation method is compared with the evaluation method based on neural network and the evaluation method based on AHP. The assessment test was carried out in A university, and the physiological data of the whole students were collected to establish a large medical database. A total of 900 students were randomly divided into 300 groups to investigate the psychological stress of all the students objectively and quantitatively. The objective investigation results are used as reference, the evaluation results of the three methods are compared with the quantized data, and the evaluation error of the method is obtained. The accuracy of evaluation is judged by analyzing the error of evaluation.

6.2 Test Result

Table 1 below is the statistical results of relative error of psychological stress assessment for students in A university by using three psychological stress assessment methods. Each group of 30 people was used to calculate the mean value of relative error in the group, and the results were compared.

Analyzing the data in Table 1, we can see that the relative error of the whole method is less than 10% when we evaluate the psychological stress of randomly grouped students. But the relative error of AHP based psychological stress assessment method fluctuates between 22–36%. The relative error of the method based on neural network fluctuates

Table 1 Comparison of relative errors of psychological stress assessment methods/%

Group	Psychological stress assessment method in this paper	Psychological stress assessment method based on neural network	Psychological stress assessment method based on AHP
1	7.46	21.31	24.38
2	9.15	20.02	23.72
3	7.23	16.24	23.20
4	5.96	19.63	33.84
5	6.58	14.46	35.55
6	9.72	20.47	23.39
7	9.07	21.59	23.27
8	8.41	15.60	25.62
9	8.74	15.28	30.58
10	9.39	15.05	29.93

between 14% and 23%. It shows that the evaluation result of this method is more close to the objective evaluation result and the evaluation result is more reliable.

To further validate the method assessing the performance of psychological stress in college students. Here, the F1 value is used as the experimental index, and the comparative experimental results are shown in Fig. 3.

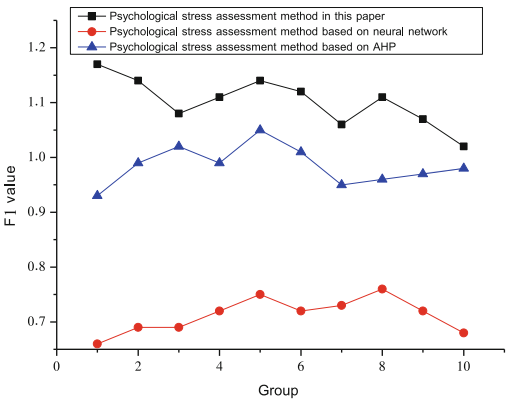

Fig. 3. Comparison of F1 value of college students' psychological assessment

From Fig. 3, the F1 value of this method is higher than the other two methods, on average, higher than 1.0. To sum up, this paper puts forward a method of college students' psychological stress automatic assessment based on medical big data, which can simplify the assessment process. And the use of big data analysis and other theories to improve the accuracy of psychological stress assessment.

7 Conclusion

The psychological pressure of students themselves is increasing, which seriously affects the normal study and life. This paper studies the automatic evaluation method of college students' psychological stress based on medical big data. First, the crawler technology was used to crawl psychologically related medical data. Then pre-process the medical data. Normalize the value interval of the data, and generally take the projection space [0, 1] as the projection space. Physiological indicators are used as specific quantitative standards to evaluate students' psychological stress state. Through data mining, the relationship between physiological data and psychological stress scale was established. So far, the automatic assessment of college students' psychological stress has been realized. Compared with other evaluation methods, the results of the proposed method are closer to the real results, and the evaluation accuracy is higher. Using this method to evaluate the psychological stress of college students can help tutors identify students who need more attention, reduce workload, and detect abnormal students early.

References

1. Li, C.: College graduate employment under the impact of COVID-19: employment pressure, psychological stress and employment choices. Educ. Res. **41**(07), 4–16 (2020)
2. Kalra, P., Sharma, V.: Mental stress assessment using PPG signal a deep neural network approach. IETE J. Res., 1–7 (2020)
3. Sun, Y., Yang, J.: Assessment of psychological pressure based on BSTL and XGDT. J. Quant. Econ. **37**(04), 148–158 (2020)
4. Yue, G.-x., Liu, J.-h., Liu, F.: Medical big data filling and classification simulation based on decision tree algorithm. Comput. Simul. **38**(01), 451–454+459 (2021)
5. Ding, Y., Chen, X., Fu, Q., et al.: A depression recognition method for college students using deep integrated support vector algorithm. IEEE Access **8**, 75616–75629 (2020)
6. Zhang, S., Ji, X.-Q., Yang, G., et al.: Research on the method of evaluating psychological stress by combination of ECG and EEG. J. Changchun Univ. Sci. Technol. **43**(02), 127–134 (2020)
7. Jia, J.: A study on the psychological pressure sources of college students and the countermeasures of psychological intervention. J. Pingdingshan Univ. **36**(06), 119–123 (2021)
8. Jin, X., Zhang, L.: Research on employment psychological pressure and coping ability improvement of college graduates from the perspective of psychological resilience. J. Qiannan Normal Univ. Nationalities **41**(04), 86–90 (2021)
9. Parthiban, K., Pandey, D., Pandey, B.K.: Impact of SARS-CoV-2 in online education, predicting and contrasting mental stress of young students: a machine learning approach. Augmented Hum. Res. **6**(1), 1–7 (2021)
10. Chenghui, B., Xinyu, W., Wenxin, G., et al.: Investigation and analysis of psychological stress of clinical medical degree postgraduates. China Continuing Med. Educ. **14**(09), 102–106 (2022)