Influence Model of Analyzing the Effect of Mental Health Level Based on Big Data Mining System

Xiaolong Wang

Faculty of Teacher Education, Pingdingshan University, South Weilai Road, Xinhua District, Pingdingshan 467000, Henan, China

Correspondence should be addressed to Xiaolong Wang; wangxl@pdsu.edu.cn

Received 21 February 2022; Revised 21 April 2022; Accepted 3 May 2022; Published 28 June 2022

1. Introduction

Accompanied by many psychological problems, especially contemporary students [1, 2], when freshmen enter the university campus, they will face a completely different living environment from the past [3]. It is difficult for many students to adapt to this new environment. Moreover, great changes have taken place in the interpersonal relationship around me [4]. I need to make new friends, leave my parents for the first time, and deal with everything by myself. Freshmen will have many negative psychological emotions, such as depression, anxiety, loneliness, etc., resulting in no interest in learning, unwillingness to communicate with others, etc. [5]. At present, universities have to conduct psychological investigation on students when freshmen enter the university, which has also accumulated a large amount of psychological data. However, how to use these psychological measurement data to get more meaningful results, so as to better carry out psychological education? Data mining technology includes many algorithms: cluster analysis (or unsupervised learning), association rule mining, prediction, time series mining, and deviation analysis. We will select appropriate algorithms according to the characteristics of psychological data to achieve the expected data
mining objectives, as shown in Figure 1, data analysis, and mining technology [6]. With the large-scale enrollment expansion of universities, the psychological problems of students are more prominent. Psychological education has been paid attention to, and the research on psychological problems is imminent. Among the psychological problems of students, anxiety and depression have become important risk factors affecting students’ physical and mental health. Only by their healthy growth and talent can we ensure that the cause of socialism with Chinese characteristics has successors and prosperity.

2. Literature Review

To solve this research problem, Tang et al. proposed tan algorithm (seminaive Bayesian algorithm). Tan relaxed the assumption of conditional independence between attributes and made the results of naive Bayesian algorithm into a tree structure, allowing each node to rely on at most one node other than the parent node [7]. Sp-tan (seminaive Bayesian algorithm) proposed by Li et al. is another tan algorithm. Sp-tan adopts greedy heuristic search algorithm. When selecting each edge, it will select the edge that will improve the accuracy of the whole classifier the most [8]. Tabbakha and Razavi combine inert learning with tan, which also weakens the hypothesis of conditional independence [9]. Jie et al. made a corresponding comparison between LBR (rule extraction algorithm) and tan and proposed lazy tan (seminaive Bayesian algorithm) in combination with their characteristics. The quality of a classifier cannot be fully evaluated by the accuracy of classification [10]. Liu et al. used AUC as the classification degree to add parent nodes [11]. Usui et al. combined boosting with tan and proposed a higher performance classification algorithm [12]. Alharbi and Shahrjerdi proposed another two-layer improved algorithm, which divides the attribute set into strong attribute set and weak attribute set. Any two attributes in the strong attribute set have dependencies, and the weak attribute sets are conditionally independent. The result of Bayesian network is a probability graph model, which also has a good structure and does not require conditional independence between attributes [13]. Kumar et al. applied Bayesian network to the analysis of primary liver cirrhosis and tested its hypothesis with a confidence of 95% [14]. Sharma et al. used Bayesian belief criterion to establish Bayesian network model in incomplete data. Someone proposed smart BN, which can be effectively used to predict human actions in video and can dynamically change the number of nodes and the relationship between nodes [15]. Lukyanov et al. assume that all points of a given cluster obey the same probability distribution, and the objects in the data set are determined according to the maximum probability value in the distribution. Hierarchical clustering, also known as agglomerative clustering, is also a very organized clustering technology. Hierarchical clustering is a method based on greedy algorithm. Each time, it calculates the similarity of data points, then selects the closest elements to form a class, and then inserts them into the original data set. The end condition of iteration is that there is only one point in the data set [16]. Through continuous test and analysis, it is found that the main symptom affecting students’ mental health is obsessive-compulsive disorder. Viewing the model with obsessive-compulsive disorder as the classification target, we can understand that anxiety disorder and interpersonal relationship also play a great role. Set the target attributes as anxiety and interpersonal relationship and the output variables as the remaining 9 factor variables to mine the main causes of obsessive-compulsive disorder, so as to provide reference for staff guiding mental health.

3. Method

3.1. Main Process of Data Mining. After years of exploration and research, people have summarized the basic process of data mining technology [17, 18]. It includes cleaning, extracting, and transforming the required data from the initial data that has not been cleaned up, generating a data set, establishing a classification or clustering model on this, and finally extracting and analyzing the information [19] (the specific process is shown in Figure 2).

3.2. Decision Tree. Decision tree is a very classic classification algorithm, which has good classification effect, and its result model has good interpretation function. Decision tree is a tree data structure composed of decision nodes and decision leaves. A leaf node can determine the category of an instance, and the function of the node is to determine how to select the next node in the test case by comparing the attribute values. For discrete attribute a, there are h possible values from A = d1, . . . , A = dh. For continuous attributes, each node has a country value. You can judge which branch should be selected by comparing with the threshold. In fact, the classification process of the decision tree is a process of moving the instance from the tree root to the leaf node. The class marks owned by all leaf nodes of the instance are the class marks owned by the instance. At present, the commonly used decision tree algorithms include ID3, C4.5, cart, etc. (as shown in Table 1). The construction algorithm of decision tree is similar. It is a construction method based on greedy thought. The division of nodes is obtained by calculating the information moisture, but the algorithm adopts different information moisture calculation methods [20].
3.3. Principle of Decision Tree. The construction of decision tree is recursively realized by top-down greedy algorithm. In each internal node, select the test attribute with the best classification effect to classify the training sample set, and recursively call the process to construct the following sub-branches until all attributes are used or all training samples belong to the same category. If the data instance and the node type in the decision tree are the same, it will be classified into the same class. If the two are different, the instance is placed as a new node in the corresponding decision tree. Repeatedly, the decision tree containing only one root node can be extended to a complete decision tree [21], as shown in Figure 3.

3.4. Data Acquisition. The data used in this study comes from the SCL-90 psychological data of first-year students in a university. There are 1643 people in this test, 989 girls and 654 boys [22]. The data mining process of students’ mental health evaluation is shown in Figure 4.

3.5. Data Preprocessing. Data preprocessing is an important link in the process of data mining. Data mining usually deals with data containing a lot of noise, fuzzy data, redundant data, or incomplete data. In the mental health evaluation data of students, incomplete data and invalid data are caused by students’ carelessness or other reasons, which will lead to a lot of inaccurate noise data. Due to the existence of these worthless data, it will eventually affect the accuracy of mining analysis results. Through data preprocessing, the level of mining can be greatly improved and the time spent in analysis can be reduced [23–25].
3.5.1. Data Selection. Data selection is a common data processing method for data analysis and mining in the early stage. It is the first step of data preprocessing. Due to the large scale of the original data set, mining and analyzing all data sets cost a lot of operation resources and operation cycle, so it is necessary to select data from the data set to reduce the impact on the results [26]. According to the mining project objectives, collecting and finding the information records in the data set can not only simplify the data content, but also find the internal relations between attributes and the laws hidden behind the data. Delete the useless information of the student basic information table, including student ID, ID number, name, date of birth, native place, telephone number, and other attributes. This information will only affect the efficiency of mining calculation. For the attribute of students' nationality, because the students in school are mainly Han and there are few other nationalities, the deletion of nationalities has no impact on the mining results.

Delete the useless information in the “student mental health evaluation form,” including the student number, gender, department, major, and other attributes of students, with the selection score of 90 questions in the psychological evaluation symptom checklist SCL-90, and retain 10 psychological dimension factors as the analysis content of data mining [27].

Finally, the data fields associated with the mining task are determined by deleting the useless attribute information in the above two tables. The data set required for the student basic information table is composed of gender (XB), household registration (HK), only child (DSZN), and family status (JTZZ). The data required by the student mental health evaluation form are composed of obsessive-compulsive symptoms (QPZZ), depression (YY), somatization (QTH), hostility (DD), anxiety (JL), interpersonal sensitivity (RJGX), psychosis (JSBX), phobia (KB), paranoia (PZ), and others (QT).

3.5.2. Data Cleaning. The main purpose of this operation is to eliminate redundancy, errors, and noise in data. Data cleaning is mainly to filter and remove duplicate data, supplement and improve incomplete data, and correct or delete wrong data. Duplicate data is mainly information with the same attribute value, incomplete data is mainly missing due information, and wrong data is mainly information written directly to the database without judgment.

3.5.3. Data Integration. Data integration is the process of integrating records from multiple related data sets into a new one to the mining target content. The data used in the paper mainly comes from the student basic information table and SCL-90 mental health evaluation table. The two tables are connected through the associated field XH (student number), and a new student mental health evaluation table is generated from the data set determined in the “data selection” process, as shown in Tables 2 and 3.

3.5.4. Data Specification. Data specification is a crucial link in data mining processing [28–30]. In data processing, we must first convert the data into a data form in line with data mining. The conversion principle usually uses continuous data discretization and discrete data classification. In this paper, the data standard operation is carried out for the information of “student SCL-90 mental health evaluation form.” The main processes are as follows.

Data discretization: the continuous data discretization of mental health test scale is helpful to data mining operation. According to each item in the symptom checklist SCL-90, grade 1 to 5 scores were taken, and 10 factors reflected psychological symptoms. If any of these factors scores more than 2 points, the screening can be regarded as positive. Therefore, the 10 factor scores of psychological symptoms are divided into two intervals: symptomatic and asymptomatic, of which more than 2 points are symptomatic and less than 2 points are asymptomatic.

Data categorization: there are many attribute values of student household registration and family economic situation. Classification conversion is required before data mining. Finally, the household registration is divided into rural (HK1) and urban (HK2), and the family economic situation is divided into difficult families (JT1) and non-difficult families (JT2).

xb1 and XB2 are used to represent men and women in gender; BX1 represents the department of nursing, bx2 represents the department of pharmacy, BX3 represents the department of medical technology, bx4 represents the department of clinical medicine, and bx5 represents the Ministry of Public Affairs; DS1 and DS2 are used to indicate whether they are only children.

All attributes in the student mental health evaluation form have passed. The codes after the above principles and specifications are shown in Tables 4 and 5.

The data table of each attribute in the student mental health evaluation table after data standardization is shown in Tables 6 and 7.

3.6. Constructing Decision Tree

3.6.1. Basic Strategy of Decision Tree Induction. Firstly, the splitting criterion of the algorithm is used to find an attribute as the splitting attribute of the training sample set. Then recursively call the above method for the subset on each branch to establish the branch on the node. With the growth of the tree, the training sample set is recursively divided into smaller and smaller subsets until all subsets contain only samples of the same category; that is, they reach the leaf node. Finally, a decision tree classification model similar to flowchart is generated.

3.6.2. Attribute Selection Metrics. Let the data be divided into D as a training sample set containing class marks, and the class label attribute has m different values, then it is defined as m different classes $C_i (i = 1, 2, \ldots, m)$, $C_{i,D}$ is the sample set of class $C_i$ in $D$, $|D|$ is the number of samples in $D$, and $|C_{i,D}|$ is the number of samples in $C_{i,D}$. 
3.6.3. Information Gain. Let node \( n \) store all samples of data division \( D \). The expected information required for sample classification in \( D \) is given by the following formula:

\[
\text{Info}(D) = -\sum_{i=1}^{m} P_i \log_2(P_i),
\]

where \( P_i \) is the probability that any sample in \( D \) belongs to \( C_i \), which is calculated by \( |C_i \cap D|/|D| \). In fact, the above formula is only the proportion of the number of samples of each class in the total number of samples. \( \text{Info}(D) \) is also called the entropy of \( D \). Entropy is a statistic used to measure the degree of chaos of a system.

Suppose that the samples in \( D \) are divided according to attribute \( a \), and attribute \( a \) has \( V \) different values \( \{a_1, a_2, \ldots, a_V\} \). If the value of attribute \( a \) is a discrete value, attribute \( a \) can divide \( D \) into \( v \) subset \( \{D_1, D_2, \ldots, D_v\} \), where the value of the sample in \( D_j (j = 1, 2, 3 \ldots v) \) on attribute \( A \) is \( a_j \). These subsets correspond to each branch growing from node \( N \). The expected information required for the sample classification of \( D \) based on attribute \( a \) can be obtained from the following formula:

\[
\text{Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j),
\]

where \( |D_j|/|D| \) is the weight of the subset whose value is \( a_j \) on attribute \( A \). \( \text{Info}_A(D) \) is the expected information required to classify \( D \) samples based on attribute \( A \).

Knowing the value of attribute \( a \) leads to the reduction of entropy, which can be obtained from the following formula:

\[
\text{Gain}(a, D) = \text{Info}(D) - \text{Info}_A(D).
\]
3.6.4. Gain Rate. Based on the splitting of attribute XH (student number), because everyone’s student number is different, there will be as many divisions as the number of student number attribute values, and these divisions are pure, and each division has only one data record. According to formula (2), the expected information required for the division of D samples according to XH (student number) can be obtained:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).$$  

(3)
samples can be divided into two categories: xb0 and XB1. There are 370 samples in XB0, of which 40 samples have a value of 1330 on attribute mg and 0 on attribute mg. There are 590 samples in xb1, of which 230 samples have a value of 1360 on attribute mg and 0 on attribute MG. According to formula (2), classify the samples in D as

\[ \text{Info}_{XB}(D) = 0.78334879. \]  

(8)

Therefore, according to formula (3), the information gain of attribute XB (gender) can be obtained as

\[ \text{Gain}(XB) = \text{Info}(D) - \text{Info}_{XB}(D), \]

(9)

\[ \text{Gain}(XB) = 0.07379965. \]

Then, according to formula (6), the splitting information of attribute XB (gender) can be obtained as follows:

\[ \text{SplitInfo}_{XB}(D) = -\frac{370}{960} \times \log_2 \left( \frac{370}{960} \right) - \frac{590}{960} \times \log_2 \left( \frac{590}{960} \right) \]

\[ \text{SplitInfo}_{XB}(D) = 0.96177796. \]  

(10)

Finally, according to formula (5), gain rate of attribute is as follows:

\[ \text{GainRatio}(XB) = \frac{\text{Gain}(XB)}{\text{SplitInfo}_{XB}(D)}, \]

(11)

\[ \text{GainRatio}(XB) = 0.07673252. \]

Using the same method, the information gain rates of attributes ZY (Major), SY (place of origin), DS (only child or not), DQ (single parent family or not), and JJ (family economic status) are calculated as follows:

\[ \text{GainRatio}(ZY) = 0.02068808, \]
\[ \text{GainRatio}(SY) = 0.03070646, \]
\[ \text{GainRatio}(DS) = 0.18323913, \]
\[ \text{GainRatio}(DQ) = 0.12276085, \]
\[ \text{GainRatio}(JJ) = 0.11974200. \]

(12)

The sample is divided into two subsets according to whether it is only child or not. Repeat the above steps to classify the sub-data set of each branch, and then export the branch again. With the increase and extension of branches, the sample data set is recursively divided into smaller sub-data sets.

### 4. Results and Analysis

#### 4.1. Construct the Decision Tree of Students’ Psychological Problems

Steps 1 and 2 are performed recursively on each of the split sub-data sets.

The class label attribute Mg (interpersonal sensitivity) has two different values: 1 (symptomatic) and 0 (asymptomatic). Therefore, the training sample set has two different categories. We first calculate the expected information of training sample set D classification as follows:

\[ \text{Info}(D) = -\frac{270}{960} \times \log_2 \left( \frac{270}{960} \right) - \frac{690}{960} \times \log_2 \left( \frac{690}{960} \right) \]

\[ \text{Info}(D) = 0.85714844. \]  

(7)

Next, the expected information of each split attribute needs to be calculated. Taking XB (gender) as an example, attribute XB has two different values: xb0 (male) and xb1 (female). Therefore, according to the value of attribute XB,
Through continuous testing and analysis, the main symptom affecting students’ mental health is obsessive-compulsive disorder. Therefore, when viewing the model with obsessive-compulsive disorder as the classification target, the results shown in Figure 5 can be obtained, according to C4.5 algorithm principle. It can be seen from Figure 2 that anxiety disorder and interpersonal relationship also play a great role.

Set the target attribute as anxiety level and interpersonal relationship level, respectively, set the output variable as the remaining 9 factor variables, and execute the data flow. The results are shown in Figures 6 and 7, respectively.

Dig out the main causes of OCD, as shown in Figure 8.

4.2. Analysis. From various angles, on the whole, the psychological quality of students is healthy. In this application, the factor of compulsion in students’ psychology occupies a relatively high proportion. Anxiety, interpersonal relationship, and paranoia have a great impact on goal attribute obsessive-compulsive disorder.

It can be seen from Figure 2 that if the anxiety level = medium, there is a tendency of obsessive-compulsive disorder regardless of the degree of interpersonal relationship. If anxiety level = none, when paranoia level = [mild, moderate], obsessive-compulsive symptom level = light; when paranoia level = none, it is related to interpersonal relationship and hostility; if paranoia level = “heavy” and “extremely heavy”, obsessive-compulsive symptom level = none. If the degree of anxiety = light, there is a tendency of obsessive-compulsive disorder regardless of the degree of interpersonal relationship. If the degree of anxiety = severe, the degree of obsessive-compulsive symptoms = moderate.

As can be seen from Figure 3, if the degree of depression = medium, the degree of anxiety = medium. If the degree of depression = none, when the degree of terror = medium, the degree of anxiety = light, and when the degree of terror = [none, light, heavy], there is almost no anxiety. If the degree of depression = mild and the degree of obsessive-compulsive symptoms = none, there is no anxiety tendency. If the degree of depression = severe, the degree of anxiety = severe.

As can be seen from Figure 4, if the degree of depression = medium, the degree of interpersonal relationship = medium. If the degree of depression = none, when the degree of terror = light and there is psychosis, the degree of interpersonal relationship = light. If the degree of depression = mild and there is obsessive-compulsive disorder, there are problems in interpersonal relationship.

As can be seen from Figure 5, in the known mining results, it is found that the causes of students’ psychological obsessive-compulsive disorder are mainly distributed in family atmosphere, family structure, and origin. Children from healthy families are full of hope for life and have great confidence in their emotional life. Due to the lack of parental
care, lack of sense of security, nerve sensitivity, and emotional vulnerability, students who have both parents died are always timid in doing things, and their psychological problems are very significant. Unsound families with single parents or divorced parents will always do harm to their children’s mental health in varying degrees and levels.

5. Conclusion

This paper proposes to influence model of analyzing the effect of mental health level based on big data mining system. Through continuous test and analysis, it is found that the main symptom affecting students’ mental health is obsessive-compulsive disorder. Viewing the model with obsessive-compulsive disorder as the classification target, we can understand that anxiety disorder and interpersonal relationship also play a great role. Set the target attributes as anxiety and interpersonal relationship and the output variables as the remaining 9 factor variables to mine the main causes of obsessive-compulsive disorder, so as to provide reference for staff guiding mental health. Future association rule algorithms can be used to analyze students’ attribute data, with more intensive research.

Data Availability

The data that support the findings of this study are available from the author upon reasonable request.

Conflicts of Interest

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References


