Computer Science and Information Systems 19(3):1583–1596 https://doi.org/10.2298/CSIS210404044Y

Data Mining Technology in the Analysis of College Students' Psychological Problems

Jia Yu* and JingJing Lin

NanChang JiaoTong Institute, Nanchang, Jiangxi, China,330022
fishpanda2011@163.com

Abstract. This paper expounds on the research status of data mining and the status quo of college students' psychological health problems, deeply analyzing the feasibility of introducing data mining technology into the analysis of college students' psychological health. After studying and analyzing the decision tree technology of data mining, and taking the psychological health problem data of the students in a university in 2021 as the research object, this paper uses the decision tree to analyze the psychological health problem data. The main work includes the following: determining the mining object and mining target; preprocessing the original data; and according to the characteristics of the data used, choosing the C4.5 algorithm of the decision tree to construct the decision tree of the students. Finally, based on the analysis and comparison of the decision tree model before and after pruning, classification rules are extracted from the optimal decision tree model, thus providing a scientific decision-making basis for mental health education in colleges and universities. After comparing the classification results with the known categories in the test set, the accuracy rate was found to be 75%. Using the alternative error pruning method and test data set, the classification accuracy was 79%, and after PEP pruning was 84%.

Keywords: data mining, data preprocessing, decision tree, college students' psychological problems.

1. Introduction

Data mining, also known as knowledge discovery, is conducted to discover hidden information deposits from massive data. It is a comprehensive application of statistics, artificial intelligence, databases, and other technologies [1]. Knowledge discovery in databases (KDD) was first proposed by Dr. Gregory Piatetsky Shapiro at the 11th International Joint Conference on Artificial Intelligence (IJCAI) in August 1989. However, at that time, multimedia outlets tended to use the term "data mining" rather than KD [2]. From 1989 to 1994, the American Association for Artificial Intelligence held four international symposiums on KDD. In 1995, the American Association for Artificial Intelligence renamed the symposium as an international conference. The first conference was held in the same year, and it became an annual event. ACM established the data mining and knowledge discovery committee (SIGKDD) in 1998 [3], and began to organize the ACM SIGKDD International Conference (KDD) in 1999. The

* Corresponding author

conference is the top annual conference in the field of data mining research. In addition, there are many conferences, organizations, and academic groups focusing on data mining and knowledge discovery, among which the more famous ones are IEEE ICDM, PAKDD, SDM, FSKD, MLDM, and VLDB. Data mining technology has been successfully applied in many fields, such as finance, insurance, securities, telecommunications, transportation, and retail. Many foreign computer companies also attach great importance to the development and application of data mining systems. Typical data mining systems include Enterprise Miner, intelligent miner, set miner, Clementine, warehouse studio, see5, cover story, knowledge discovery workbench, Extra, quest, and dB miner. Data mining has become a standard term and encompasses text mining, image mining, web mining, prediction analysis, and processing massive data (big data). Google vigorously promotes web mining and text mining, and launched the powerful Google Analytics in 2006 [4], which can carry out access data statistics and analysis on a target website and provide a variety of parameters for website owners to use.

A survey showed that the situation of college students' weariness, dropping out of school, suicide, and hurting others is mostly caused by mental health problems, and the number of students with poor mental health has been on the rise. A survey of 126,000 college students across in China shows that 20.3% of them have psychological health problems, mainly manifesting as fear, anxiety, obsessive-compulsive disorder, depression, and neurasthenia [5].

College students face different psychological health problems in different stages of development and deal with them in different environments. Freshmen have some psychological problems in some university. Freshmen are in the second "weaning period" of their life. If they leave home and do not adapt to the new college environment, they are prone to the contradiction between independence and dependence. Facing the gap between the ideal and reality, they are then prone to frustration, anxiety, depression, and even neurasthenia. However, their mental health will obviously improve as they adapt to university life. Before graduation, students' psychology will also change a lot. Junior college students are ready to take the entrance examination of postgraduate, and undergraduates are ready to take the entrance examination of postgraduate. Most of the remaining people are under pressure due to difficulties in finding employment after graduation. Confusion, fear, and irritability disrupt their psychological balance and make some people depressed, lose confidence, unable to find a goal in life, and even feel that life is meaningless [6].

The present paper attempts to apply data mining technology to the analysis of college students' psychological problems, in order to excavate the useful knowledge hidden in the data. We then hope to use this knowledge to predict the mental health status of college students more accurately, so as to provide a scientific basis for the planning and decision-making of mental health education. An additional goal is to make mental health education more effective [7].

2. Related Work

This section mainly introduces the purpose of classification, classification steps, and evaluation criteria of classification methods. We then introduce several commonly used classification algorithms, and analyze and compare the application scope, advantages, and disadvantages of these algorithms.

Data classification is a very important function in data mining and is employed to construct a classification pattern based on the given data. A model is then used to classify unknown data records in a database. At present, [x] is mainly used in commercial applications, but is gradually being applied to other fields[8].

First, the data in the training sample set is analyzed. According to the characteristics of the training samples with known class labels, a description (model) of the known class is constructed. Then, the class description (model) is used to classify the unknown data, and the class label of the unknown data is predicted.

Data classification can be summarized into two steps: building a model, and then using that model for classification.

2.1. Construction model (classifier)

We build a model (classifier) that describes a predefined data set or concept set. Each tuple in the scheduled data set has a class label; that is, it belongs to a scheduled class. The model is constructed by analyzing the attribute description of the database tuple. The data tuple set is regarded as a sample set, and the sample subset randomly selected from the sample set to construct the classification model constitutes a training sample set, in which each tuple is a training sample. Because the class label of each training sample is known in advance, the learning process in constructing the model is a guided learning process. Usually, the first step learning is used to construct the model in the form of decision tree, formula, or rule. Then, this model can be used to classify other samples and provide a deeper understanding of the database.[9]

2.2. Using a model (classifier) for classification

First, the prediction accuracy of the model should be evaluated. In classification method, samples independent of training samples and with known class labels are randomly selected as test samples; that is, the test data set is completely different from the training data set. Each test sample in the test data set is used to learn the prediction class according to the classification model and compare it with the known class label. If the labels are the same, the classification is successful. The accuracy of the model refers to the percentage of the number of test samples correctly classified by the model. If the accuracy is acceptable, the model can be used to classify the data tuples with unknown class labels.

3. Application of data mining in the analysis of college students' psychological problems

In Section 2, a decision tree algorithm was selected as the data mining algorithm to be employed for the analysis of college students' psychological problems. In this chapter, we use the decision tree algorithm to create a decision tree model of college students' psychological problems, prune it, and then extract rules from it. These rules are then used to find those attributes that have a greater impact on psychological problems. We then analyze their impact degree so as to provide a decision-making basis for psychological counselors of college students. The final decision tree model is also used to predict new data so as to aid in the prevention of psychological problems in college students.

3.1. Determining the research object and mining target

In this study, the mining data are the basic personal information of students provided by the psychological counseling center of a university and 2012 survey results of university students (the survey employs the symptom checklist 90). This paper randomly selects a psychological symptom and establishes a decision tree model through classification mining to determine whether college students have the psychological symptom.[10]

3.2. Data acquisition

In order to obtain the required data, this paper uses symptom checklist 90 (SCL-90) (see Appendix I) to evaluate 1700 students in 42 classes, 29 majors, and 5 departments in a university. The age of the students tested ranged from 17 to 24 years. A total of 1700 self-rating scales were distributed, and 1640 were actually recovered. According to SCL-90 instructions (see Appendix II), the scores of each student's psychological factors were calculated.

The data needed for the analysis of psychological problems are managed by the database management system SQL Server 2008. We create a "psychology" database under the graphical interface of SQL Server Management Studio of SQL Server 2008. In this "psychology" database, we then create a "personal psychological problems" table, and a "personal basic information" table. Table 1 presents the definition of the table structure of the "personal psychological problems" table. Table 2 defines the table structure of the "personal basic information" table. The "personal psychological problems" table stores the scores of all the psychological factors of all the tested students. The scores of each psychological factor can be used to judge whether the students have symptoms related to nine kinds of psychological problems. The data of the "personal basic information" table is directly obtained from the psychological counseling center of the University.

Table 1. Definition of table structure of "personal psychological problems"

Field name	type	maximum length	meaning
XH	C	10	student number
QT	n	4	somatization
QP	n	4	forced
Mg	n	4	interpersonal sensitivity
YY	n	4	depression
JL	n	4	Anxiety
Hostile	n	4	DD
KB	n	4	terror
PZ	n	4	paranoia
JS	n	4	psychotic

Table 2. Table structure definition of "personal basic information" table

Field name	type	maximum length	meaning
XH	C	10	student number
XM	C	10	name
XB	C	2	gender
CS	d	8	date of birth
MZ	C	10	ethnic groups
ZY	C	10	major
Source of SY	C	20	students
Is DS	C	2	the only child
Is DQ	C	2	a single parent family
JJ	C	10	family economic status

3.3. Data preprocessing

Data cleaning. Since some students do not actively participate in responding to SCL-90, or are distracted or hindered by some external factors, not all surveys were returned, and some were irregularly filled in. Thus, this data can only be used after cleaning. The specific method employed in this paper is as follows. First, the non-standard filling items in the self-assessment scale are regarded as vacant items, and then the number of vacant items is counted. If there are no more than three vacant items, the most selected item in the item will be used as the items option. Otherwise, the students' self-assessment scale will be voided directly by the method of ignoring samples. After data cleaning, 200 self-rating scales were removed, and 1440 were retained for later use. Since the calculation result of each self-rating scale corresponds to one record in the "personal psychological problems" table, the number of records in the "personal psychological problems" table is 1440.

Data integration. Data integration is the task of integrating the related data according to the mining goal and build a new and closely coupled data set. The data used in this paper is related to students' "personal basic situation" table and "personal psychological problems" table. Through XH (student number), the tables of "personal basic information" and "personal psychological problems" are naturally linked to form a table of "basic information and psychological problems".

Data selection. In order to improve the efficiency of mining, it is necessary to further simplify the data set, that is, delete those useless data according to the mining task, and determine the data set to actually be used in mining. The data selection work in this paper proceeds as follows:

1. Directly delete XH (student number), XM (name), and CS (date of birth) in the "basic information and psychological problems" table because these attributes are meaningless for mining.

2. The MZ (nationality) in the table of "basic information and psychological problems" is deleted because the minority students account for only 1.9% of all students (most of them are of Han nationality, and this attribute has little effect on the mining results).

3. Since the specific task of this paper is to analyze which basic attributes of students are related to interpersonal sensitivity symptoms and how relevant they are, the Mg (interpersonal sensitivity) attribute in the "basic situation and psychological problems" table is selected for data mining.

Finally, in the table of "basic situation and psychological problems", there are seven attributes describing the basic situation of students, namely, XB (gender), ZY (major), sy (place of origin), DS (only child or not), DQ (single parent family or not), JJ (family economic status), and Mg (interpersonal sensitivity). The table of "psychological problem analysis" is used as the data set to construct the decision tree model of whether students have interpersonal sensitivity symptoms.

Data conversion. Since some of the data in the data set is not conducive to data mining, before mining, we should carry out data conversion according to the principle of continuous data discretization. In this paper, the data conversion of the "psychological problem analysis" table includes the following steps:

1. The attribute value of Mg (interpersonal sensitivity) is continuous, and it is discretized. According to the symptom Checklist-90 instructions for use, "when using the 0-4 level scoring method, which psychological factor score is greater than or equal to 1, indicating that the person has symptoms in which psychological problems." Because this paper adopts the 0–4 grade scoring method, the value of the psychological factor is either 1 (symptomatic) or 0 (asymptomatic). In this way, the continuous data is discretized, which is convenient for mining.

2. Although the values of some attributes are discrete, there are too many categories, and thus the data must be converted before mining. For example, there are 29 kinds of attribute values of majors. We classify majors into three categories, namely, literature and history, science and engineering, and art. Student places of origin are scattered across the country, and there are bound to be many categories of attribute values. We

classify the attribute value of students' place of origin into three categories: urban, county, and rural. In this way, the number of data categories is reduced, which is conducive for mining.

3. In order to improve the efficiency of data mining, we try not to use Chinese characters. Chinese attribute values are replaced by English characters, numbers, or a combination of English characters and numbers. For example, xb0 represents male, zy1 represents literature and history, etc.

3.4. Constructing the decision tree

Basic strategy for decision tree construction. Most decision tree algorithms are greedy algorithms. Based on the training sample set and the associated class labels, the decision tree is constructed by top-down recursive partitioning [5]. First, an attribute is found as the splitting attribute of the training sample set using the splitting criterion of the algorithm. This attribute is taken as the root node of the tree, and the branches are established according to the different values of the attribute. The training sample set is divided into subsets. This whole method is called recursively for each branch to establish new nodes and branches. As the tree grows, the training sample set is recursively divided into smaller and smaller subsets until all subsets contain only samples of the same class; that is, they are all leaf nodes. Finally, a decision tree classification model similar to a flow chart is generated, and it can be used to classify new samples. A path from the top root node to the bottom leaf node is called a classification rule.

Attribute selection metrics. Attribute selection metrics, also known as split rules, determine how to split samples on a given node. Here are two popular attribute selection metrics: information gain and gain rate.

Let d be the training sample set with class label, and the class label attribute has m different values. There are m different classes, which is the sample set of classes in D. | D | is the number of samples in D, and | is the number of samples in D.

1**. Information gain.** Let node n store all samples of data partition D. The expected information for the sample classification in D is given by the following formula:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 \left(p_i \right) \qquad (1)$$

where [x] is the probability that any sample in D belongs to and is calculated by | / | D |. In fact, the above formula only uses the proportion of the number of samples of each class to the total number of samples. This is also known as the entropy of D. Entropy is a statistic used to measure the degree of chaos in a system.

Suppose the samples in D are divided by attribute a, which has different values. If the value of attribute a is a discrete value, then the attribute a can divide d into subsets, where the value of the sample in attribute a is a discrete value. These subsets correspond to the branches growing from node n. The expected information for sample

classification based on attribute a to attribute d can be obtained by the following formula:

$$Info_A(D) = \sum_{j=1}^{y} \frac{|D_j|}{|D|} Info\left(D_j\right) \tag{2}$$

where [x] is the weight of the subset with a value based on attribute a.
The expected information needed for the classification of samples divided into D.

Knowing the value of attribute a results in a decrease in entropy, which can be obtained from the following formula:

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

Classification allows us to extract information from a system and reduce the degree of confusion in that system; this results in the system becoming more regular, orderly, and organized. The more chaotic the system, the greater the entropy. It is obvious that the optimal splitting scheme is the one with the largest entropy reduction. Therefore, the ID3 algorithm selects the attribute with the maximum information gain (a) as the splitting attribute on the node.GainAN.

2. Gain rate. The ID3 algorithm uses information gain as the attribute selection metric and is biased to select attributes with more attribute values. Consider the following example. Based on the split of attribute XH (student number), i.e., because everyone's student number is different, there will be as many partitions as the number of student number attribute values; these partitions are pure, and each partition has only one data record. According to formula (2), we can obtain the expected information of the sample according to XH (student number): (d) = 0 XH. According to formula (3), the information gain of this attribute is the largest, and it will be used as the first splitting attribute. However, for classification, it is meaningless to divide based on student number. C4.5 and ID3 have the same basic principle, but the difference is that in order to compensate for ID3's disadvantage of using information gain to select attributes with more values, C4.5 uses gain rate instead as the attribute selection metric (splitting rule). The definition of information gain rate is as follows [48]:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{4}$$

In the above formula, split information is used to normalize the information gain:

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{r} \frac{|D_j|}{|D|} \log 2\left(\frac{|D_j|}{|D|}\right) \tag{5}$$

SplitInfo(D) represents the information generated by dividing the training sample set into plans corresponding to the outputs of the attribute test.ADAvv

Constructing the decision tree of college students' psychological problems. The data set used in this paper has three attribute values of specialty, student origin, and family economic status; meanwhile, the other attributes have only two attribute values. If the ID3 algorithm is used, it will tend to select attributes with more attribute values when selecting split attributes, which will affect the quality of mining. Conversely, the C4.5 algorithms can solve this problem well while still inheriting all the advantages of the ID3 algorithm. In order to make the mining results more accurate, we choose to use the C4.5 algorithms to construct our decision tree.

In this paper, we randomly selected 2/3 of the data set (960 records) to comprise the training sample set for decision tree mining, and the other 1/3 (480 records) as the test sample set for decision tree testing. Below we describe the whole process of building the decision tree model using the C4.5 algorithms.

1. Using formula (1), formula (2), formula (3), formula (4), and formula (5), the information gain rate of each split attribute in the training sample set is calculated.

2. By comparing the information gain rate of each split attribute, the split attribute with the maximum information gain rate is set as the root node of the decision tree. If the attribute has several values, the data set is split into several sub data sets. If the attribute has only one value, the split ends.

3. Step 1 and step 2 are performed recursively on each of the split sub data sets.

The training sample set is randomly selected from the data, we can know that 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. There are 270 samples corresponding to class value 1, and 690 samples corresponding to class value 0.

According to formula (1), we first calculate the expected information of training sample set classification:

$$Info(D) = -\frac{270}{960}\log 2\frac{270}{960} - \frac{690}{960}\log 2\frac{690}{960} = 0.85714844$$

Next, we need to calculate the expected information of each split attribute. For example, XB (gender) has two different values: xb0 (male) and xb1 (female). Therefore, xb0 and xb0 can be divided into two categories. There are 370 samples in xb0, of which 40 samples have 1330 values in attribute mg, and 40 samples have 0 values in attribute mg. There are 590 samples in xb1, among which 230 samples have 1360 values in attribute mg, and 230 samples have 0 values in attribute mg.

Meanwhile, it can be seen that DS (only child or not) has the largest information gain rate of the attributes. Taking DS (only child or not) as the root node leads to two branches: only child and non-only child. The samples are thus divided according to these branches. We repeat the above steps to classify the sub data set of each branch, and then create new branches. With the increase in the number of branches, the sample data set is recursively divided into smaller sub data sets. Finally, the decision tree model is generated.

3.5. Tree pruning

Pruning is an effective method to simplify a decision tree. During the construction of the decision tree, the algorithm does not consider the existence of noise and outliers in data. In this way, many branches are constructed based on anomalies in the sample set, which leads to overfitting between the generated decision tree and the training samples. Furthermore, the generated decision tree will be overgrown, which reduces its practicability and readability and increases its dependence on the sample data. Moreover, while the accuracy of the decision tree may be very high, when the decision tree is used for the classification of new data, the accuracy can become very low. In order to deal with the problem of overfitting data and to obtain more general classification rules, it is necessary to prune the decision tree. In this paper, the decision tree has 50 leaf nodes, from which 50 classification rules can be extracted. This is a large number of rules, and some of them are too long to understand.

Pruning is carried out to improve the readability and classification accuracy of the decision tree model.

Pre-pruning. First pruning is to prune a tree by stopping its construction in advance. Because this pruning is done before node splitting, it is called pruning first. If the tree splitting ends at a node, the node will be replaced by a leaf, and there may be several samples belonging to other categories in the training sample set covered by this node [6]. Therefore, information gain, statistical significance, Gini index, and substitution error rate can be used to evaluate the pros and cons of splitting. If the partition of a node results in a split lower than a predetermined threshold, the partitioning is stopped. However, it is often difficult to determine the appropriate threshold. If the threshold value is too high, the decision tree will be too simple and lose its application value; if the threshold value is too low, the tree will be pruned too little, and thus some redundant branches cannot be cut off. The specific method of pruning first also includes presetting the minimum number of training samples contained in the training sample subset of each node, stopping the node splitting when it is less than the predetermined value, and presetting the highest level of the decision tree to limit its growth.

1. The principle of alternative error pruning.

In the process of constructing the decision tree, when the samples contained in the branches of a given node belong to the same class or there are no remaining attributes that can be used to continue to split the sample set, the division of the decision tree is stopped. If the subset of a branch of the decision tree is further split during the calculation, and if the number of samples in different classes varies greatly, the formula for error substitution rate can be introduced:

$$\text{pure}\frac{n-n'}{m} \tag{6}$$

where [x] denotes the number of samples in the branch, [x] denotes the number of samples of most categories on the branch, and [x] represents the total number of samples in the training sample set. If the substitution error rate calculated by the formula is lower than the predetermined threshold, the branch will be transformed into a leaf node and

identified by most categories on the branch so as to prevent overfitting to a certain extent.

2. Pruning decision tree based on substitution error rate

In this paper, before constructing the decision tree model, we set the threshold at 0.3%. While building the tree, the substitution error rate of each branch is calculated according to formula (6). If the substitution error rate is lower than the predetermined threshold, the branch is cut and replaced with a leaf node, and then the category corresponding to most of the samples in the cut branch is used as the class label of the leaf node. For example, when node 4T in Figure 4.3 splits based on XB (gender), the value of XB (gender) is 70 on the branch with xb0 (male), and is 7068, the total number of known samples is 960. According to formula (6), the substitution error rate of the branch can be calculated as 0.2%, which is lower than the predetermined threshold of 0.3%. Then the division is stopped and the branch is replaced by a leaf node. The class label of the leaf node is considered to be category 0 (asymptomatic). The substitution error pruning method is employed when using C4.5 to generate the decision tree model.

Post-pruning. The post-pruning algorithm first conducts fitting and then simplification. First, the algorithm constructs a decision tree model that completely fits the training sample set, and then deletes several subtrees and replaces them with leaf nodes. Next, the class of most of the training samples in the subtree is used to identify the class of the leaf node. Since post-pruning is based on all the information of the decision tree and conducted according to a certain standard, the effect of post-pruning is generally better than that of first pruning. The four most commonly used post-pruning algorithms are reduced error pruning (Rep), probabilistic error pruning (PEP), minimum error pruning (MEP), and cost complexity pruning (CCP). When choosing the pruning algorithm, we should balance the accuracy and complexity of the decision tree, combine the characteristics of different algorithms with the actual sample set, and choose one algorithm or a combination of several algorithms according to the specific situation. Generally speaking, the rep algorithm is the simplest, but it is not suitable for a sample set with small amount of data; PEP is one of the most accurate algorithms; the CCP algorithm gets a smaller tree size than the rep algorithm; and the MEP algorithm gets a larger tree size but has a lower precision than the PEP algorithm.

1. Selection basis of the PEP pruning algorithm

We choose the PEP algorithm for pruning due to the following two considerations:

(1) The PEP algorithm does not need a pruning set, and this is useful when pruning a decision tree model based on a small data set. Since the data set used in this paper is small, the PEP algorithm is a good choice.

(2) The PEP algorithm has high pruning efficiency. Because this paper is to study the practical application of data mining technology in the analysis of college students' psychological problems, rather than the research of data mining algorithm, so the work efficiency is higher. Furthermore, the PEP algorithm adopts a top-down strategy, which makes it faster and more accurate than other pruning algorithms.

2. Principle of the PEP pruning algorithm

The PEP pruning algorithm was proposed by Quinlan in order to overcome the shortcoming that the rep algorithm needs to prune a data set separately. It does not need

pruning set, but prunes based on the error estimation of training data set. However, this will lead to a large error in the estimation error rate. Therefore, Quinlan introduced a continuity correction. In other words, it is assumed that each leaf node automatically misclassifies 1/2 of the instances it covers, and a constant is added to the training error of the subtree. When calculating the standard error rate, the continuous correction follows a binomial distribution.

We use [x] to represent the original tree; [x] represents the subtree with the node as the root node; (T) represents the number of misclassified instances at the node; and (T) represents the number of all instances covered at the node.

The classification error rate of node T is

$$r(t) = e(t) / n(t) \tag{7}$$

The PEP algorithm modifies this to

$$r'(t) = [e(t) + 1/2] / n(t) \tag{8}$$

Let s be the subtree TT of T, and l (s) denote the number of leaf nodes of TT. Then, the classification error rate of TT is

$$r'(T_t) = \frac{\sum_s [e(S) + 1/2]}{\sum_s n(S)} = \frac{\sum e(S) + \frac{L(S)}{2}}{\sum n(S)} \tag{9}$$

For the sake of simplicity, the error rate is replaced by the number of errors:

$$e'(t) = e(t) + 1/2 \tag{10}$$

The PEP algorithm adopts a top-down method. If the calculation result of a non-leaf node is

$$e'(T_t) \le e'(T_t) + SE\left[e'(T_t)\right] \tag{12}$$

then t is cut off and replaced by its corresponding leaf. The standard error is defined as

$$SE\left[e'(T_t)\right] = \sqrt{\frac{e'(T_t)\left[n(t) - e'(T_t)\right]}{n(t)}} \tag{13}$$

Because each subtree is visited at most once during pruning, the PEP algorithm is faster and better than other algorithms and is recognized as one of the most accurate pruning algorithms. However, the top-down pruning strategy of PEP algorithm will bring the same "field of view effect" as the first pruning method, that is, when a node is pruned according to the pruning conditions, the node may have subordinate nodes that should not be pruned according to the pruning conditions.

1. Pruning the decision tree with PEP

The PEP algorithm is used to calculate each non-leaf node in a top-down way. This is done according to formulae (10), (11), and (13). The pruning results of each non-leaf node are shown in Table 4.5. The calculation results of non-leaf nodes, T24, T27, T29, T29, T32, t33, can make equation (12) hold. Therefore, the subtree with the six non-leaf nodes as root nodes is cut and replaced with leaf nodes. Then, the class label of the leaf node is identified by the class of most of the samples in the subtree. Finally, we obtain a pruned decision tree model.

Obviously, the size of the pruned decision tree model is smaller than that of the directly generated decision tree model, and this improves the readability of the decision tree and the classification speed of new data.

Result analysis. From the generated rules, it can be seen that the most relevant attribute of students with interpersonal sensitivity symptoms is whether they are the only child. After analysis, we can draw the following conclusions: among the students with interpersonal sensitivity symptoms, the proportion of those who are an only child is significantly higher than those who are a non-only child; the proportion of female students is significantly higher than that of male students; the proportion of art majors is higher than that of other majors; and the proportion of students from single parent families is higher than that of students from non-single parent families. Among the only child attribute, female students from rural areas or majoring in art generally have symptoms. In addition, it is more common for students with poor family economic status to have symptoms, and most of them come from rural areas. Therefore, when carrying out psychological counseling and counseling related to interpersonal sensitivity, we should pay special attention to high-risk groups, such as students who are an only child, females majoring in art, students from poor families, and students from single parent families, and give them timely and appropriate guidance and help. These rules obtained through data mining technology can help university management departments firmly grasp the initiative of work, practice the guiding ideology of prevention first and treatment second, and carry out targeted psychological counseling.

4. Model accuracy evaluation

This study randomly selected 2/3 of 1440 records (960 records) as the training sample set; the remaining 1/3 (480 records) was used to form the test data set. It was known whether each student had interpersonal sensitivity symptoms in the test data set. After comparing the classification results with the known categories in the test set, the accuracy rate was found to be 75%. Using the alternative error pruning method and test data set, the classification accuracy was 79%, and after PEP pruning was 84%. Obviously, the classification accuracy of the decision tree model generated after pruning by PEP was the highest. This was followed by the decision tree model generated by introducing the substitution error first pruning method, and lastly the decision tree model generated directly. The pruned decision tree model obviously has more accurate classification than the directly generated decision tree model. Overall, the decision tree model constructed in this paper achieved the expected classification effect.

5. Conclusion

According to the requirements of decision-making analysis of mental health education for college students, the classification and mining of college students' psychological problems is fully realized. First, we determine the mining objects and data mining objectives. Based on SCL-90 results and basic information of students, a decision tree model ofwhether students have interpersonal sensitivity symptoms is constructed. Then the training sample set is obtained by preprocessing the data. According to the characteristics of the training sample set, the C4.5 algorithm is selected to construct the

decision tree model and prune it. Then, the classification rules are extracted from the decision tree model and analyzed. Finally, the accuracy of the model is evaluated. This paper also compares the original tree with the pruned tree in terms of scale, extracted classification rules, and classification accuracy. The conclusion is that the pruned decision tree model is simpler, easier to understand, and more efficient than the directly generated decision tree.

Acknowledgment. This work was supported by Humanities and Social Science Research Projects in Colleges and Universities in Jiangxi Province in 2020：”A Study on the Relationship between Psychological Capital, Academic Burnout and Subjective Well-being of Independent College Students”SZZX20138.

References

1. Zhao Ding. Research on emotional data analysis and psychological early warning system based on data mining and intelligent computing [J]. Electronic production, 2021 (02): 88-90
2. Xu Meijuan, Xiao Xiong, xiong Ye, Dong Liangliang. Analysis of the "heart" way out of data mining technology in college mental health education [J]. Science and technology and innovation, 2020 (22): 43-44 + 47
3. Dai Qun, Wang Yibo. Research on academic early warning system of college students based on data mining [J]. Rural staff, 2020 (20): 230-231
4. Liang Yueying, Yuan Fang, Zhang Jian. Analysis of research hotspots of health education for college students at home and abroad based on data mining [J]. Chinese Journal of multimedia and network teaching (first ten issues), 2020 (10): 46-48
5. Yang Xun. Research on early warning of College Students' Psychological Crisis Based on big data technology [J]. Digital world, 2020 (10): 88-89
6. Chang Ni, Yu Yanjun. Dilemma and countermeasure analysis of psychological crisis early warning in Higher Vocational Colleges from the perspective of big data [J]. Intelligence, 2020 (26): 92-94
7. Gu Yongcheng. Analysis and Research on students' mental health based on Internet behavior data [D]. Guangdong University of technology, 2020
8. Fu ronghua. Application of association algorithm based on decision tree in rural college students' information system [J]. Hubei Agricultural Sciences, 2020,59 (10): 150-153 + 158
9. Chen haoquan, Hu Ruiyu, Zhao Zheng, Wang Shi. Design of mental health problem prevention platform based on data mining and database [J]. Science and technology horizon, 2020 (14): 49-51
10. Lin Jingyi, Li Dakun, Wu Pingxin, Wang Xu, Zhou Yan. Modeling and analysis of mental health early warning based on social data mining [J]. Electronic technology and software engineering, 2020 (08): 172-173

Jia Yu has Master of Applied Psychology, and works as Lecturer. She graduated from the Jiangxi Normal University in 2012 and worked in NanChang JiaoTong Institute. Her research interests include mental health of college students and counseling.

JingJing Lin has Master of Management. She graduated from the Jiangxi University of Finance and Economics in 2011. She worked in NanChang JiaoTong Institute. Her research interests include government behavior and public policy.

Received: April 04, 2021; Accepted: May 20, 2022.