Application of Data Mining in Cultivation of Student Mental Health

Huanchang Qin¹, Zaiqun Wu¹, Meisuo Wang² *

¹College of Information Engineering, Baise University, Baise, Guangxi, China, 533000
²School Clinic of Baise University, Baise, Guangxi, China, 533000

*Corresponding author E-mail: 13097765363@163.com

Abstract. Under the education concept of new era, colleges and universities need to focus on cultivation of students' mental health besides the necessary quality education. Effective planning and decision-making of mental health education help students establish correct outlook on life and values, with great significance on their comprehensive development. With the rapid development of science and technology, data mining technology has been gradually applied to cultivation of students' mental health. The work focused on the application of data mining in mental health education of students. Through rational application of data mining technology, the current models of mental health education for college students were further improved to provide reference for future mental health education.

Keywords: Data mining, Mental health, Application

1. Introduction
High-speed social and economic development has led to increasingly fierce social competition and complex interpersonal relationships. Psychological pressure of people is continually increased in such a social environment. The survey shows that people will experience great changes in emotions and awareness under complex social environment. Especially for college students, psychological problems are easily caused if they are not effectively guided. The healthy growth of students will be negatively influenced to seriously affect harmonious development of the society[1]. Therefore, psychological health of college students has been focused by all sectors of the society in recent years. Effective improvement of mental health of college students has also become a major issue for the development of colleges and universities. Research on students' psychological problems cannot be separated from the support of a large amount of data. Through the analysis on the large data of psychological problems, the laws and patterns behind data can be summarized to provide scientific and effective guidance for mental health education. Data mining technology meets this demand, which has been widely used in cultivation of students' mental health.
2. Data Mining and Knowledge Discovery

2.1. Definition of Data Mining
Data mining mainly refers to the process searching for hidden information among a large amount of data using specific algorithms. As an information search technology, data mining relies on computer and network technologies. Necessary computer science and technology is involved in the process of information search. Meanwhile, methods of online processing analysis, information retrieval, expert system and pattern recognition are used to achieve rapidity and accuracy of information search. Data mining technology has wider application compared with other technologies of data information search. Structured or unstructured data can be searched with data mining, with high application value.

2.2. Data Mining Technology
Data mining technology is an information technology generated on the concept of data mining, which is the result of long-term research and development of database technology. Regarding various types of data as the management and research objects, scientific and complete database and data warehouse are established with such technology through unified organization and management of data. On the base of original data warehouse, multiple operations can be performed on the data, such as addition, deleting, modification, processing, analysis, report, and printing [2]. In addition, the data warehouse can clean up the inside data. Different problems of the data are solved according to actual situation, thus promoting development of data mining. Therefore, the effectiveness of data warehouse technology is closely related to data mining technology.

2.3. Association Rule Mining Algorithm—Apriori Algorithm
Decision Tree Algorithm and Apriori Algorithm are two main algorithms in data mining technology. Wherein, Decision Tree Algorithm is a classification algorithm with the advantage of fast modeling speed. Attributes of continuous value and discrete value can be processed, with intuitive result for easy understanding and acceptance. The association rule mining algorithm is mainly used to discover underlying internal relationship among the data. It aims to dig out the implicit internal relations in a large amount of data information, excavating the original unfamiliar knowledge. This algorithm is widely used in transactional data analysis [3]. The advantage of Apriori Algorithm is that the scientficity and effectiveness of decision making can be enhanced through accurate predictions with visualized and clear presentation.

Based on the above advantage analysis and the studies on student mental health education, the work selected data mining algorithm, where Apriori Algorithm is mainly focused. During the research on students' psychological problems, we attribute the cause of a problem to the occurrence of another issue. That is, the occurrence of one thing causes another thing. Such kind of connected and symbiotic situation is the association rule in psychological problems of students.

2.4. Data Mining Process
Data mining involves links of object determination, data preparation, data mining, result analysis and knowledge assimilation. The effective convergence of each link constitutes a complete and systematic data mining process. In short, data mining technology is used for analysis and processing on a large amount of related data in the data warehouse. Finally, effective, practical information is discovered as a basis to find relevant laws, making scientific and effective decisions [4]. Figure 1 shows specific data mining process.
Figure 1. Specific data mining process
Specific contents of each link in data mining process are described as follows:

2.4.1 Determine the object and purpose of data mining. As the first step of data mining, determining purpose can ensure the pertinence of data mining. Though the results are unpredictable, the targeted problems need to have predictability. Expected prediction effect cannot be obtained without specific purpose for data mining. In cultivation of students' mental health, the focus is the existing problem in implementation of students' mental health education. After determining the direction of data mining direction, we can conduct targeted data mining work.

2.4.2 Data preparation. This link mainly includes data selection, preprocessing and transforming. Wherein, data selection refers to the searching of internal and external data related to students' mental health training. The searched information amount should be enough to support this study, allowing selection of the data information that can be excavate in a large amount of data information. After data selection, data preprocessing is performed for better data analysis. In general, preprocessing includes integration, transformation, reduction, and compression of data information. Meanwhile, it is necessary to determine the type of mining operations. As the last link, data transforming aims to convert the data into an analysis model established for the mining algorithm. The scientificity of the established analysis model is directly related to success of data mining, with great importance.

2.4.3 Data mining. The converted data is excavated with data mining technology. Apriori Algorithm is used for data mining to obtain correlation between different data.

2.4.4 Results analysis. This link refers to interpretation and evaluation of the results. The analysis method depends on excavation operation, where visualization technology is the most frequent method of current mining results analysis.

2.4.5 Knowledge assimilation. The knowledge gained from analysis is integrated into organizational structure of the business information system[5].

3. Application of Data Mining in Cultivation of Students' Mental Health

3.1. Overall Demand
Currently, the problem of students’ mental health cultivation has aroused attention of education departments. On-campus psychological counseling center and related management system have been established by colleges and universities according to their actual development. In mental health management system, teachers can have comprehensive and in-depth understanding of students' mental health problems based on data information. Meanwhile, mental health problems can be analyzed and classified into different types. However, there are not functions of data query and screening in traditional mental health management system, which is difficult to demonstrate the relevance among different various types of mental health problems. In contrast, data mining technology includes functions of data querying and retrieving, allowing the discovery of required information in large
amount of data. Therefore, data mining technology was applied to the study of students' mental health training in the work.

3.2. Data Preparation

Data preparation mainly includes following four aspects:

3.2.1 Data Selection. To ensure pertinence and reference of data mining results, relevant data was selected as mining data from mental health status test of students entering Baise University in 2013. The students majored in Chinese Language and Literature, Social Work, Electronic Information Technology, Life Science Technology and Computer Network Technology. Questionnaire was used for data acquisition. A total of 525 questionnaires were distributed, where 518 were recovered with 494 valid questionnaires. Tables 1 and 2 show the main structure of mental health test data:

**Table 1. Basic information of students**

| Student No. | Name | Gender | ID No. | Nationality | Only child | Birth date | Student type | Grade | Major | Home place |
|-------------|------|--------|-------|------------|------------|------------|--------------|-------|-------|------------|

**Table 2. Data of students’ mental health status**

| S | OC | IS | D | A | H | PA | PI | P | Others | Result |
|---|----|----|---|---|---|----|----|---|--------|--------|

* S: Somatization; OC: Obsessive-compulsive; IS: Interpersonal sensitivity; D: Depression; A: Anxiety; H: Hostility; PA: Phobic anxiety; PI: Paranoid ideation; P: Psychoticism

3.2.2 Data Preprocessing The fundamental purpose of data mining is to obtain valuable data for research in a large amount of data information[6]. However, too much data information in the database easily leads to interference of useless data in mining process, such as vacant data, noise data and inconsistent data. Thus, subsequent analysis and excavation work will be negatively affected, even bringing impact on the entire data mining results. The results of mining data cannot meet the requirements without high-quality data. Therefore, it is important to preprocess the data.

Three items are involved in data preprocessing:

a. Data extraction. Not all data is suitable for mining data throughout the data mining process. In the large amount of data information, some data have no effect on mining results, while some can greatly affect mining efficiency. Such data will waste lots of time on data mining, even leading to fault of the results. For current data in student mental health database, some data information of attributes has no effect on mining results, such as student number, name, nationality, ID number, etc. However, such data will increase mining time, without improvement for the results. Therefore, above attribute values were removed during data mining to improve mining efficiency. Then, the data attributes closely related to mining task in the database were selected according to characteristics and purpose of students' mental health cultivation. The attributes included gender, major, home location and “if only child?” Research indicates that such data attributes have great influence on common psychological problems of low self-esteem, anxiety and depression. After determining data attributes according to mining purpose, data extraction was performed to ensure that the mined data satisfies requirements of mining task.

b. Data cleaning. Manual and automatic methods are involved in data cleaning. Wherein, manual method requires lots of time and costs in verification of data consistency, which is only used for
small-scale data cleaning. Automatic cleaning is used for medium or large-scale data. The following two aspects need to be completed in automatic cleaning process: On the one hand, the data is defined to determine types of data error; on the other hand, the wrong instance should be corrected with targeted measures after they have been searched and identified. Manual methods can be used in this study due to small amount of data. A preliminary screening was implemented to ensure data validity while recovering the questionnaire, eliminating the test data with great missing values. Therefore, the task for data cleaning in the work was to confirm consistency and integrity of the data.

The main factor leading to incomplete and inconsistent data was nonstandard and wrong filling of students. Incomplete and inconsistent data result in a large amount of noise data, null data and duplicate data, which are the objects of data cleaning. Null data was the most important cleaning object. Generally, deletion will not affect mining results in the data set with small number of null values. However, when the proportion of null values reaches a certain level, deletion of data set will cause great loss of mining data, thereby affecting mining results. Therefore, null value is a very important part of data cleaning. There are different ways to make up null values. For example, the mean value in a noun variable or a group of data can be used as a substitute value for calculation. Besides, null values can be filled according to distribution pattern of the data. For example, the ratio of “only child” to “non-only child” is 5:3, so that the corresponding vacancy field can be randomly filled according to the proportion. After data cleaning, quality of the mined data was significantly improved to enhance the effectiveness and accuracy of data mining.

3.2.3 Data Conversion. As the key content of data preprocessing, the principle of data conversion is discretization of continuous data and classification of discrete data. Continuous and discrete data are two types of data information. Wherein, discrete data has the problem of too much categories. For example, attribute values of major and home place belong to discrete data, which will increase difficulty of data mining. Data conversion should be performed to minimize the impact of such attribute values. In addition, conversion of data format is of great importance, as specification of data format can effectively improve mining efficiency. Table 3 shows part of code after data conversion:

| Attribute | Attribute value | Code |
|-----------|----------------|------|
| Gender    | Male           | XB1  |
|           | Female         | XB2  |
| Only child| Yes            | DS1  |
|           | No             | DS2  |
| Major     | Chinese Language and Literature | ZY1 |
|           | Social work    | ZY2  |
|           | Electronic Information Technology | ZY3 |
|           | Life Sciences and Biotechnology | ZY4 |
| Home place| Small and medium-sized cities | JT1 |
|           | Urban          | JT2  |
|           | Rural          | JT3  |
| Somatization | With symptoms (factor score≥3) | QT1 |
|           | Without symptoms (factor score<3) | QT2 |
| Obsessive-compulsive | With symptoms (factor score≥3) | QP1 |
|           | Without symptoms (factor score<3) | QP2 |
| Depression | With symptoms (factor score≥3) | RJ1  |
|           | Without symptoms (factor score<3) | RJ2  |

Table 3. Code after conversion of student psychological mining data
3.3. Discovery of Students' Mental Health Problems Based on Apriori Algorithm

3.3.1 Data Item and Data Item Set Let \( I=\{i_1,i_2,\ldots,i_m\} \) be a set composed of \( m \) items, then each item \( i \) \((K=1, 2, \ldots, M)\) is called a data item. Wherein, \( I \) and \( m \) are data item set and its length, respectively; a data item set with length of \( K \) is called \( K \)-item set. A data item is equivalent to a field in the database, while a data item set is equivalent to a collection set composed of all fields. Let \( D \) be a multiple set of affairs. Each affair \( T \) is a set of items in data item set \( I \), namely \( TiI \). The whole affairs constitute an affair set \( D \). An affair is equivalent to the record in the database, consisting of several fields.

3.3.2 Association Rule as well as its Support Degree and Confidence An association rule is a logical implication of the form \( X \rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \), and \( X \cap Y = \emptyset \). In a data set \( D \), if \( C \% \) of the affairs containing \( X \) also contain \( Y \), then confidence degree of the association rule \( X \rightarrow Y \) is \( C \% \). If \( S \% \) of the affairs in \( D \) contain \( X \) and \( Y \), support degree of the association rule \( X \rightarrow Y \) is \( S \% \). Specific description is shown as follows:

\[
\text{Support degree: support}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} (4.1)
\]

\[
\text{Confidence: confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} 
(\text{support}(X) \neq 0) (4.2)
\]

Among association rules, support and confidence are important parameters reflecting significance and reliability of association rules. Generally, the user will set a minimum support threshold according to data mining conditions. If \( \text{support}(X) \) is greater than or equal to this threshold, then \( X \) is called frequent item-set; otherwise it is called infrequent item-set. The mining task of association rules is to generate association rules with all support and confidence that are not less than the user-specified minimum thresholds.

3.3.3 Discovery of Students' Mental Health Symptoms Based on Apriori Algorithm To better improve results of data mining, data information of the original database was divided into 10 sub-databases according to attribute values, including individual attribute values and a psychological symptom in data information. Table 4 shows the database of mental health status of some students in terms of anxiety, wherein JL1 and JL2 represent “with anxiety symptoms” and “without anxiety symptoms”, respectively.

### Table 4. Affair database of mental health status

| No. | Gender | Only child | Major | Home place | Anxiety |
|-----|--------|------------|-------|------------|---------|
| 1   | XB1    | DS1        | ZY2   | JT2        | JL1     |
| 2   | XB2    | DS2        | ZY2   | JT3        | JL2     |
| 3   | XB1    | DS2        | ZY1   | JT1        | JL2     |
| 4   | XB1    | DS1        | ZY2   | JT2        | JL1     |
| 5   | XB1    | DS2        | ZY1   | JT1        | JL2     |
| 6   | XB2    | DS1        | ZY1   | JT3        | JL2     |
| 7   | XB2    | DS1        | ZY1   | JT1        | JL2     |
| 8   | XB1    | DS2        | ZY2   | JT1        | JL1     |
| 9   | XB1    | DS2        | ZY2   | JT2        | JL2     |
| 10  | XB1    | DS2        | ZY1   | JT2        | JL2     |
| 11  | XB1    | DS1        | ZY1   | JT2        | JL1     |
| 12  | XB1    | DS1        | ZY2   | JT1        | JL1     |
| 13  | XB2    | DS1        | ZY2   | JT1        | JL2     |
| 14  | XB1    | DS2        | ZY2   | JT1        | JL2     |
| 15  | XB1    | DS2        | ZY1   | JT2        | JL1     |
| 16  | XB2    | DS1        | ZY1   | JT1        | JL1     |
| 17  | XB2    | DS2        | ZY2   | JT2        | JL1     |
| 18  | XB1    | DS2        | ZY1   | JT2        | JL2     |
| 19  | XB1    | DS2        | ZY1   | JT3        | JL2     |
Mining of association rules was performed on 30 records collected from Table 4. Figure 4 shows the diagram of Apriori Algorithm:

**Figure 2.** Apriori Algorithm

Minimum support and confidence were set as 0.2 and 0.3, respectively, discovering anxiety symptoms of students. By comparing the results with the information in Figure 2, it was found that anxiety symptoms of students were related to various factors including “if only child?” and home place, besides gender and major. For data mining of other symptoms using Apriori Algorithm, we also found that there were differences among students of different majors in terms of mental health symptoms. For example, students majoring in Chinese Language and Literature and Social Work had better performance in interpersonal sensitivity, compared with those of Computer Network Technology and Electronic Information Technology. Students of Life Science and Technology were more likely to be impulsive, with obvious hostility.

There were significant differences between rural and urban students in terms of obsessive-compulsive and anxiety symptoms. Compared with urban students, rural students live in tougher living environment and habits, leading to greater psychological pressure, interpersonal sensitivity and inferiority symptoms. No significant differences existed between male and female students in symptoms of obsessive compulsive and psychoticism. However, male students had worse symptoms of depression and paranoia than males. Meanwhile, male students had higher hostility and
compulsion than females. Due to influence of regional and language environment, ethnic-minority students generally had mental health symptoms of sensitivity, anxiety, hostility and compulsion.

4. Conclusions
In conclusion, the problem of student's mental health is of great significance for personal development as well as social harmony and stability, which needs enough attention of education institutions. As the hope of social development, college student becomes the main force in construction of socialistic China. Requirements for mental health education of college students will be gradually increased under new modes of social development. Therefore, educational institutions need to make full use of data mining technology in the future. Large amount of data is analyzed to predict the trend of students' psychological problems. Then, scientific and complete educational measures should be taken to fundamentally solve psychological problems of college students, thus promoting their healthy development.

Acknowledgement
The construction project of characteristic specialty and experimental training teaching base (center) of Guangxi undergraduate universities from 2018 to 2020: One of the achievements of the characteristic specialty of "Electronic Information Engineering" in Baise University (approval number: Higher Education of Guilin Education [2018] No. 52; No. 119).

References
[1] Yu Wang, Xiaofan Lv. Application of Data Mining in Cultivation of Students' Mental Health. Computer Programming Skills & Maintenance, 2013 (04).
[2] Guangdong He. Application of Data Mining Technology in Analysis of College Students' Psychological Problems. Hebei University, 2013.
[3] Wan Chen. Application of Data Mining in Mental Health Management System of Higher Vocational Students. Zhengzhou University, 2017.
[4] Sheng Fang. Application of Data Mining Technology in Mental Health Education for Secondary Vocational Students. Hefei University of Technology, 2013.
[5] Jinyu Zhu. Effective Use of Data Mining in Prevention of Psychological Crisis in College Students. Course Education Research, 2016 (02).
[6] Shengwang Zong, Guangdong He. Application of Data Mining Technology in Mental Health Files of University Students. Fortune Times, 2015(2X).