Application of Association Rule Data Mining in Statistical Analysis of College Students' Mental Health

Hongrui Zhang

Institute of Higher Vocational Education,Hebei Chemical & Pharmaceutical College，Shijiazhuang，Hebei,050026,China

Abstract

Strengthening the application of big data technology in data analysis can effectively improve the service capability and level of relevant statistics, and provide comprehensive and reliable information support for macro decision-making and trend analysis. This paper comprehensively reviews the research status of big data technology in the field of college students' mental health at home and abroad. Combining with the characteristics of college students' mental health statistical data and the weaknesses in statistical analysis, the feasibility of using knowledge mapping technology is demonstrated. On this basis, the blood relationship graph and influence analysis among the statistical indicators of college students' mental health were constructed through the knowledge map. The application of the knowledge map of college students' mental health statistical indicators in statistical data analysis, statistical indicator identification and statistical data quality management is proposed. Specifically, based on the concept of big data, we can establish a decision analysis platform for college students' mental health. Based on the big data technology, the data mining and analysis ability can be enhanced. In addition, it can change the traditional thinking of college students' mental health statistics and strengthen the construction of statistical team.

Keywords: *statistical analysis; big data; Mental health of College Students*

I. Introduction

With the continuous development of information technology, big data has penetrated into all walks of life. In recent years, the use of data mining technology to explore the mental health condition of college students and psychological crisis intervention model has become a hot spot for scholars at home and abroad to study[1]. The main content of this study is to use two levels of statistical analysis and data mining techniques to analyze and dig out the correlations between the factors that lead to the psychological problems of college students and the main factors that affect the psychological problems of students from the survey data. The purpose of this paper is to provide a more scientific basis for decision making in terms of psychological counseling in colleges and universities, to provide treatment plans for students with mental health problems, to provide new methods for early prevention and intervention in the mental health of college students, and to make the mental health education work in colleges and universities more rational[2].

II. Theoretical knowledge

2.1 Data mining techniques

Data mining [1] is the process of discovering relationships implicit in data from large, incomplete, noisy, fuzzy, and random data, building models, and extracting information and knowledge that is potentially valuable, credible, novel, valid, and understandable to humans. Data mining is an emerging and evolving discipline that incorporates the latest technology research results in database technology, artificial intelligence, machine learning, statistics, knowledge engineering, and information retrieval. Data mining applications are very wide. Foreign research on data mining technology has achieved fruitful results, and there are relatively successful application cases in retail, finance and insurance, and medical service industries. In practical applications, the patterns are often classified into

categories such as association analysis, classification, cluster analysis, sequence analysis, and isolated point analysis according to their actual role and the tasks of data mining[3-4].

2.2 Association rule mining

Association Rule Mining (ARM) is a fruitful and active research branch in the field of data mining for discovering interesting connections hidden in large data sets.

1) Assuming that I is a set of $I = \left[i_1, i_2, i^3, \cdots, i_n \right]$, and the dataset D be the set of transactions, where each transaction T is the set of items such that T ⊆ I. Each transaction has a denotation called TID. A transaction T contains an itemset A. An association rule is a logical implication shaped as A → B when and only when A ⊆ T, where A ⊂ I, B ⊂ I, and A ∩ B = ϕ.

2) Support(A→B) = P(A∪B) = Support(A∪B) = S. It indicates the probability that the concurrent set A ∪ B of item set A and item set B occurs in all transactions D. The measure of support reflects whether the association rule is universal or not.

3) Confidence(A→B) = P(B|A) = Support(A∪B)/Support(A) = C. It is the probability that in a transaction D in which the item set A appears, the item set B also appears at the same time. The measure of confidence reflects the reliability of the association rule.
4) Lift(A→B) = P(B|A)/ P(B). It is the ratio of the confidence level to the desired confidence level. It is a useful association rule only when its value is greater than 1.

A strong rule is an association rule with both confidence and support greater than a given threshold (minimum confidence threshold and minimum support threshold). Otherwise, it is called a weak rule. Given a transaction set D, mining association rules is to generate strong rules[5-8]

.

III. Statistical analysis of college students' mental health status

3.1 Data preparation

A total of 4320 students from the class of 2011 in a university were tested using the symptom self-assessment scale SCL_90, which was developed by the Ministry of Education's "Chinese College Students' Mental Health Assessment System" project team. The tested students came from several departments, including the Department of Mathematics and Computer Science, the Department of Mechanical Engineering, and the Department of Civil Engineering. For the mental health data of college students obtained from the survey, these data need to be pre-processed before mining. The main tasks are as follows.

3.1.1 Processing null values

Attributes such as only child, student leader, place of origin, and family structure are the main attributes related to data mining, and all of them are not allowed to have null values. The Chinese College Student Mental Health Assessment System does not deal with the missing values of the above attributes. Due to the small number of vacant values, this paper uses the manual filling method to fill the vacancy using most of the attribute values.

3.1.2 Data extraction

Attributes such as name, student number, test date, and answers to each question are not relevant to mining, while 97.2% of the students tested are Han Chinese, which has no effect on the mining results. Therefore, these attributes

were removed before mining was performed to improve the mining efficiency.

3.1.3 Data specifications

In this paper, the psychological symptom values were classified into four levels: none, mild, moderate, and severe. The continuous data "monthly household income" is disaggregated and divided into low, medium, and high intervals according to less than 2000 RMB, between 2000 and 5000 RMB, and more than 5000 RMB. The discrete data "place of origin" is transformed, For example, the remote rural areas are generalized into high-level conceptual rural areas. After generalization, the place of origin is divided into large and medium cities, small towns, and rural areas.

3.2 Statistical analysis of college students' mental health data

Of the 4320 students tested, 17.85% had somatization disorder, 61.02% had obsessive-compulsive disorder, 42.43% had interpersonal sensitivity disorder, 29.26% had depression, 31.57% had anxiety, 32.75% had hostility disorder, 23.94% had phobia, 39.47% had paranoia, and 31.64% had psychosis. In this paper, two psychological disorders with high proportion, obsessive-compulsive disorder and interpersonal sensitivity disorder, were statistically analyzed in terms of gender, only child, place of origin, student leader, family structure, and monthly family income, as shown in Figures 1 to 6.

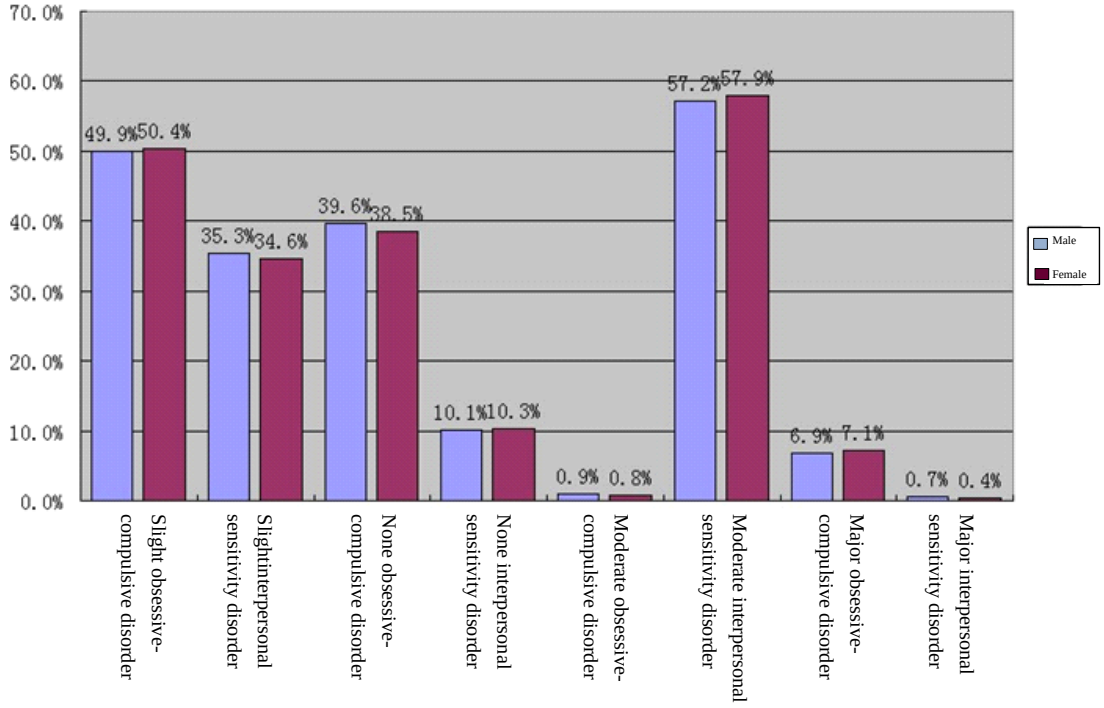


Figure 1 Graph of the relationship between gender and psychological symptoms

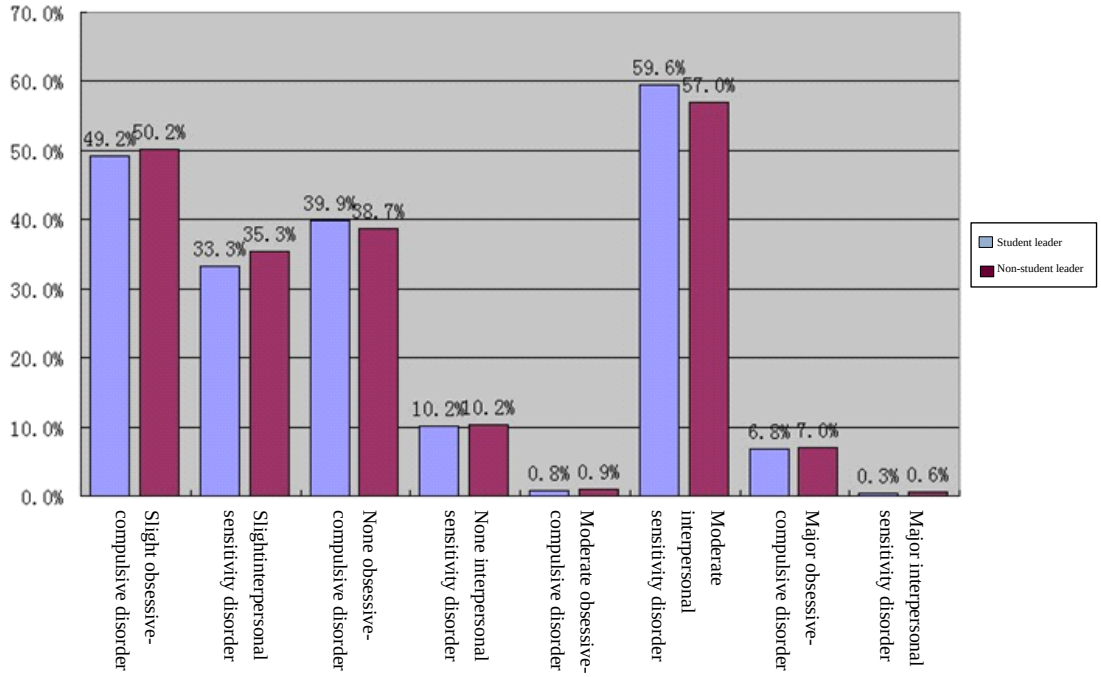


Figure 2 Graph of the relationship between student leaders and psychological symptoms

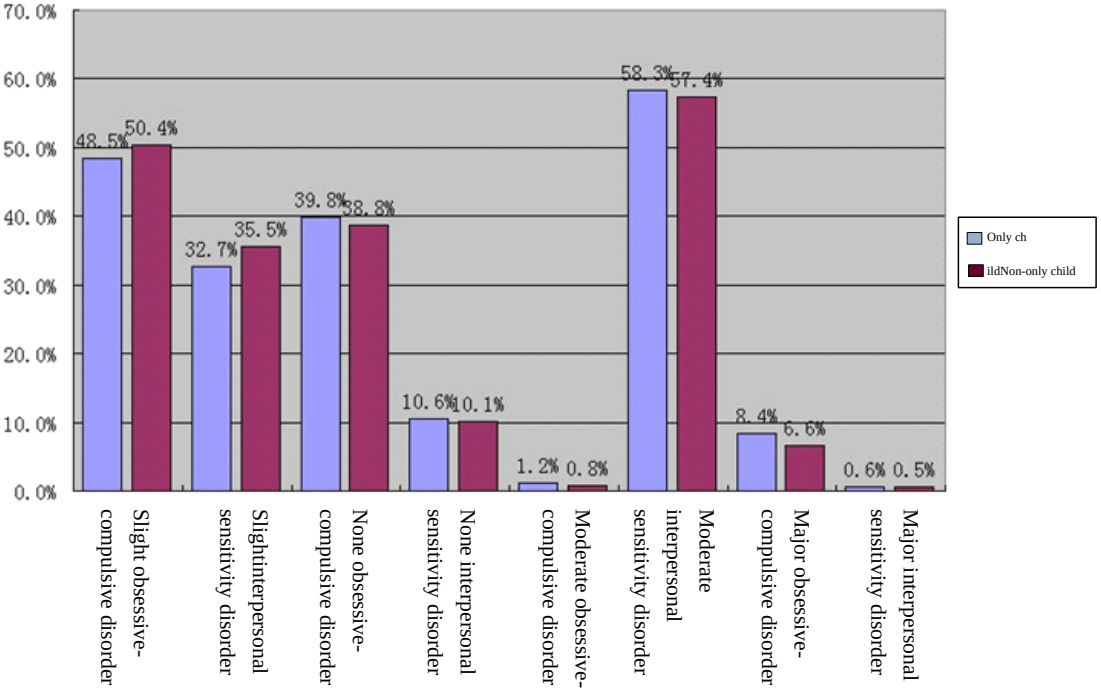


Figure 3 Graph of the relationship between only child and psychological symptoms

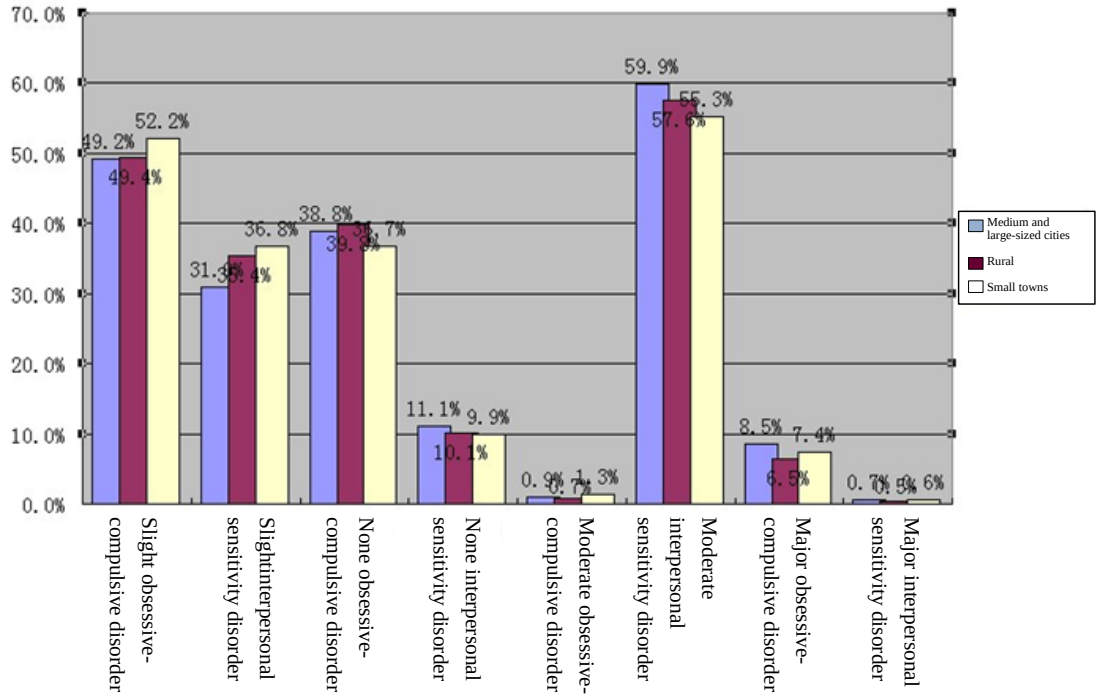


Figure 4 Graph of the relationship between place of origin and psychological symptoms

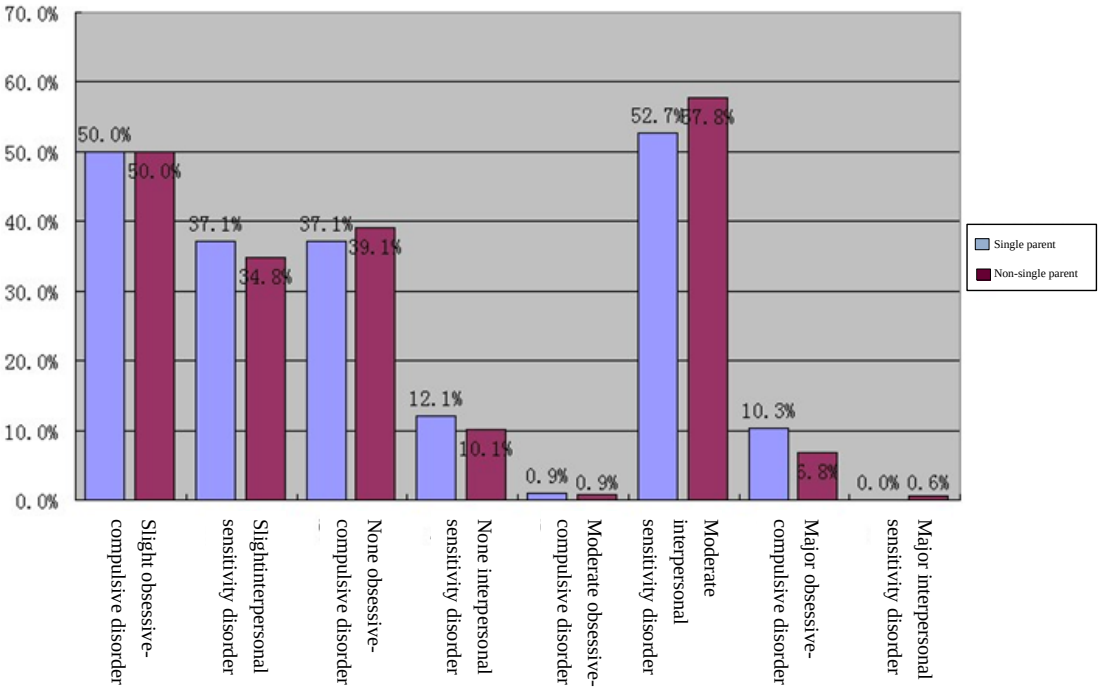
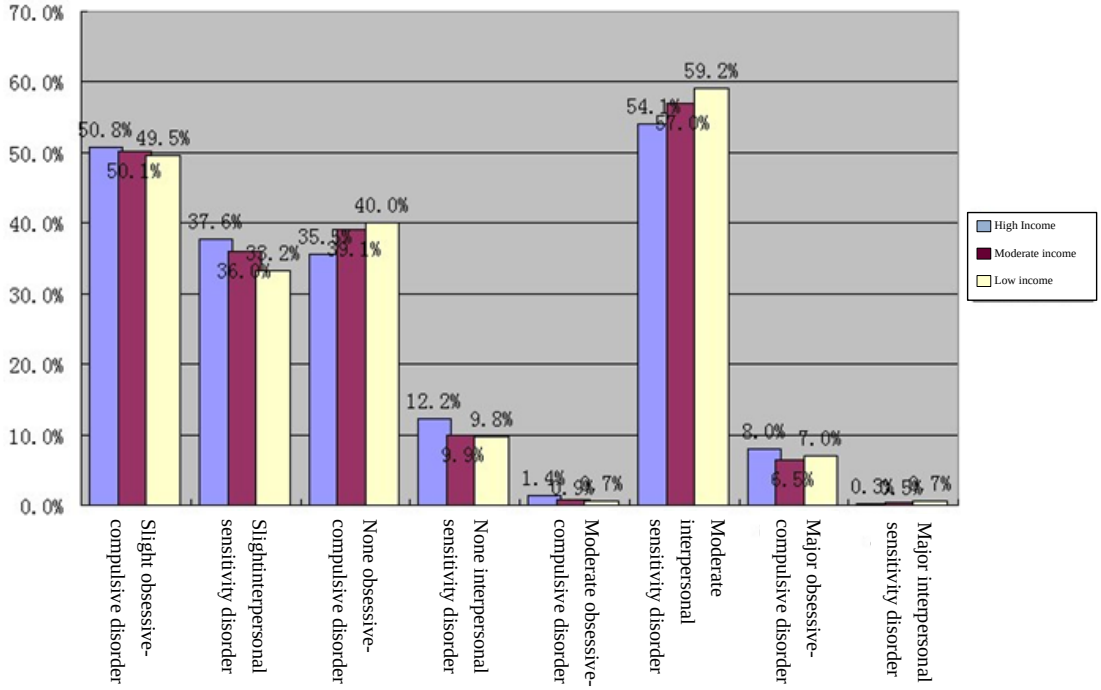


Figure 6 Graph of the relationship between household income and psychological symptoms

Figure 5 Graph of the relationship between single parent and psychological symptoms



Analysis and recommendations: Among the students surveyed, the majority of students had mild OCD and moderate interpersonal sensitivity. The proportion of female students suffering from severe OCD was higher than that of male students. Students with moderate interpersonal sensitivity were predominant, with a slightly higher proportion of female students than male students. A higher proportion of male students than female students suffered from severe interpersonal sensitivity. The proportion of student leaders without OCD was higher than that of non-student leaders. The proportion of non-student cadres suffering from mild, moderate and severe OCD was higher than that of student cadres, and the proportion of non-student cadres suffering from severe interpersonal sensitivity was twice as high as that of student cadres. Students who were only children had higher rates of moderate-to-severe OCD and moderate-to-severe interpersonal sensitivity than those who were not only children. Students in small towns had higher rates of both mild obsessive-compulsive and interpersonal sensitivity than students in large and medium-sized cities and rural areas. Students in large and medium-sized cities had a higher percentage of students without interpersonal sensitivity. Rural students had lower rates of both severe OCD and interpersonal sensitivity than students in both large and medium-sized cities and small towns. A higher percentage of students from non-single-parent families had no OCD than students from single-parent families, while a higher percentage of single-parent families had severe OCD than non-single-parent families. A higher percentage of students from high-income families were free of interpersonal sensitivity than students from low- and middle-income families. A higher percentage of students from low-income families had severe interpersonal sensitivity than students from high school income families, and a higher percentage of students from high-income families had severe OCD than students from low- and middle-income families. Students who are student leaders improve their coordination and communication skills with teachers and students in various chores of daily life, and naturally have a lower percentage of interpersonal sensitivity than non-student leaders who have interpersonal sensitivity. Only children lack interaction with siblings, grow up alone, and do not know how to handle relationships among classmates. As a result, a higher percentage of only children naturally suffer from moderate to severe obsessive-compulsive and relationship sensitivity disorders than non-only children. Children from large and medium-sized cities are influenced by their living environment, and their parents put too much pressure on them from various aspects, so the proportion of students with severe psychological symptoms is higher than that of students from rural areas. Children from single-parent families are not able to enjoy the love of their parents at the same time due to the incomplete family,

and they may have some radical views on problems and their mental health should not be neglected. Children from high-income families do not have to worry about finances, so their mental health is better than that of students from low- and middle-income families. Female students, non-student leaders, non-only children, students from small towns, single-parent families or low-income families should be given more attention by the relevant authorities and they should be provided with psychological guidance.

IV.Analysis of college students' mental health status based on association rule mining

4.1 Establishment of a multidimensional association rule mining model for college students' psychology

The current representative data mining software includes DBMin- er, Mineset, IntelligentMiner, IBM Quest and so on. In this paper, we use SPSS Clementine12.0 as the platform for mining model building and analysis, and the association rule mining models in Clementine are "GRI model", "Carma model" and "Apriori model". The algorithm can handle both Transactional and Tabular data formats [4]. In this paper, we choose the classical "Apriori" algorithm to build the model, select the direction of "both" in the type node, filter out the attributes that are not relevant to the analysis in the filter node, and build the association mining data flow, as shown in Figure 7.

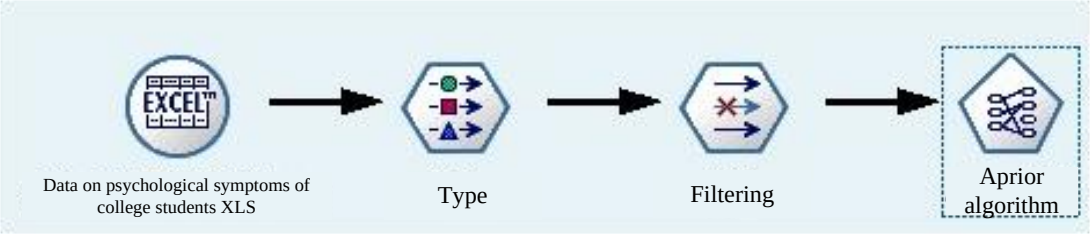


Figure 7 Multidimensional association mining data flow

4.2 Assessment and analysis of data mining results of college students' psychological symptoms

Based on the results of the statistical proportions of different attributes and different psychological symptom situations as a reference basis for the support and confidence thresholds, we continuously adjusted the support and confidence thresholds during the mining. Then, the correlations between six attributes and psychological symptoms of college students, namely, gender, student leader, only child, place of origin, family structure, and monthly family income, were obtained separately. In this paper, two psychological disorders with high proportion of obsessive-compulsive disorder and interpersonal sensitivity disorder were taken as examples for mining analysis. The mining results are shown in Tables 1 and 2.

Table 1 Association rules between attribute-OCD dimensions (Selected)

No	Association rule	Support /%	Confidence /%	Lift
1	Female, low income → light	26.898	50.602	1.013
2	Student leader, low income→None	10.741	42.241	1.084
3	Student leader, male → light	11.597	52.295	1.047
4	Male, rural→none	28.28	40.783	1.046
5	Male, only child→No	13.681	41.794	1.072
6	Small town→Light	20.139	52.184	1.045
7	Single-parent family→Light	5.185	50	1.001
8	Non-single-parent family→Light	94.815	49.951	1
9	Only child, low income→None	10.139	41.553	1.066
10	Female, non-student leader→Light	44.19	51.493	1.031

Table 2 Association rules between attribute-interpersonal sensitivity dimensions (Selected)

No	Association rule	Support /%	Confidence /%	Lift
1	Rural, high income → Light	10.185	38.409	1.1
2	Student leader, low income → None	10.741	61.853	1.074
3	Non-only child, high income → light	12.292	39.36	1.128
4	Middle income, male → Light	15.856	37.664	1.079
5	Female, low income → None	26.898	59.466	1.033
6	Student leader, non-single-parent family → None	20.116	59.839	1.039
7	Single-parent family→Light	5.185	37.054	1.061
8	Non-single-parent family, single child → None	20.324	58.884	1.023
9	Low income → None	49.028	59.16	1.028
10	Rural, male → light	28.38	35.726	1.023

Analysis and recommendations: The main manifestations of OCD among college students are excessive tension, worry, fear, and even insomnia and general discomfort during the preparation for or during examinations, which to some extent reflects the hard work of college students in their studies. Table 1 lists the degree of association between some attributes and OCD. For example, Rule 3 indicates that male student leaders accounted for 11.597% of the students surveyed, while the percentage of mild OCD among all male student leaders was 52.295%. Based on the mining results, it can be seen that the prevalence of mild OCD is higher among female students, non-student leaders, non-only children, and children from small towns, high-income families, or single-parent families, while the prevalence of no OCD is higher among male students and students from low-income families. Table 2 presents the degree of association between selected attributes and interpersonal sensitivity disorder. According to the mining results, it can be seen that the reliability of students without interpersonal sensitivity is higher among student leaders, only children, students from large and medium-sized cities, low-income families, or non-single-parent families, and there is no significant difference between male and female students. Due to the influence of the general social environment, the lack of quality education in schools, and the parents' extra care for only children, the children's selfish mentality is cultivated. Therefore, they tend to be self-centered, self-absorbed, socially utilitarian, suspicious and jealous, and righteous in interpersonal interactions and communication. Due to the dysfunctional parent-child relationship, improper parenting style of guardians, social evaluation pressure and their own weak psychological adjustment ability, children from single-parent families tend to develop insecurity, low self-esteem and lead to autism, isolation and even rebellion. Students who have served as student leaders or live in large and medium-sized cities have a wide range of social contacts and relatively rich social experience, so they handle interpersonal relationships better, while rural children are limited by their living environment, material conditions and insight, and are under too much psychological pressure.

V. Conclusion

This paper introduces data mining techniques and the theoretical basis of association rules. After pre-processing the data of college students' mental health assessment data, two methods, statistical analysis and association rule mining, are used to analyze the association relationship between students' various attributes and psychological symptoms. Based on the mining results, a deeper understanding of students' psychological problems can be obtained, and suggestions are made on how to strengthen and improve the psychological crisis intervention work of college students. In data mining, as the proportion of students with severe psychological symptoms is too low, in order to dig out the association of these symptoms, the support degree must be set very small, resulting in the generation of a lot of useless rules, making the analysis more difficult. Further research is needed regarding association rule mining.

Reference

[1] Guo Zhirong. Research on the innovation of College Students' mental health education under the background of big data [J]. Contemporary educational practice and teaching research (Electronic Journal), 2018, 000 (009): 269271

[2] Liang Juan, Luo Haiju. Application of big data mining method in college students' psychological early warning system [J]. China school health. 2018 (12)

[3] Long Shuqin. Research on big data on College Students' mental health education [J]. Journal of College of electronic engineering, 9 (3): 1

[4] Peng Jinxiang. Research on data analysis and feedback system construction of College Students' mental health under the background of big data [J]. Digital technology and application, 2019

[5] Qian Chunxia, Gu Weiwei. Research on information construction of College Students' Psychological Archives [J]. Archives and construction. 2015 (04)

[6] Shao Shuai, Liu Xuejun, Li Bin. Research on microblog emotional tendency based on key sentence analysis [J]. Computer application research. 2018 (04)

[7] Lai Lizu, Tao Rong, Ren Zhihong, Jiang Guangrong. Current situation and ethical issues of online psychological counseling in mainland China: "big data" perspective and ethical evaluation [J]. Psychological science. 2018 (05)

[8] Luo Yutong, Chen Zhibo, Tang Xing. Innovation of College Students' mental health education in the era of virtual reality technology [J]. Educational modernization. 2019 (43)