Application of Data Analysis in Mental Health Education for College Students

Ying Liu(✉)

Liaoning Jian Zhu Vocational College, Liaoning 111000, China

Abstract. ID3 algorithm is used to construct decision tree to predict the mental health status of college freshmen, so as to provide decision support for college students' mental health education. This paper introduces the main content of ID3 algorithm, discusses the data preprocessing, tree building algorithm, using decision tree to predict Freshmen mental health, and the application integration method of decision tree in MIS. The experimental results show that this method has a certain practical value in the construction of preventive mental health education mode for college students.

Keywords: Data mining · Decision tree · ID3 · Mental health education

1 Introduction

At present, colleges and universities in China have paid more and more attention to the mental health education of college students. Many schools have carried out a general survey on the mental health of freshmen, carried out relevant psychological tests on students one by one, and established students; personal mental health files on this basis. However, due to various reasons, the early warning mechanism and assistance system of mental health in many schools have not been established completely, so it is impossible to detect and intervene the students; psychological problems as soon as possible. However, due to the lack of early warning and intervention mechanism of psychological problems, students mental health diseases are often worsened, forming a vicious circle. How to establish a scientific and efficient early warning mechanism for students; mental health is a severe challenge for the current mental health education in Colleges and universities. Using data mining technology to find out the hidden information from the massive data of the existing college students; personal mental health archives database, and to provide decision support for college mental health education, will be the solution [1].

It is one of the effective ways to solve the above problems.

Data mining is a process of discovering potential, novel and valuable knowledge from a large amount of data. The tasks of data mining include: classification, clustering, association, regression, prediction, sequence analysis, deviation analysis, etc. There are many algorithms of data mining, such as decision tree, association analysis, Bayesian algorithm, neural network, genetic algorithm, rough set, fuzzy set and statistical analysis. Among them, decision tree algorithm is probably the most popular data mining

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2021, LNICST 391, pp. 9–15, 2021.
https://doi.org/10.1007/978-3-030-87900-6_2

technology. The most common data mining task solved by decision tree is classification. Compared with other algorithms, decision tree algorithm can quickly create mining model, and the created model is easy to explain. The research content of this topic is to use ID3 algorithm to build a decision tree to predict the mental health status of college freshmen, find out the most likely to have mental health problems, so that the school can achieve scientific and effective early warning, early intervention and key prevention of students; mental health problems.

2 ID3 Algorithm Principle

Decision tree is a sample based inductive learning method, which is a tree structure similar to flow chart. In the process of generating algorithm, non l eaf nodes represent attributes, while leaf nodes represent categories. The top node of the tree is the root node. A path from the root node to the leaf node forms a classification rule. Decision tree can be easily transformed into classification rules, which is a very intuitive representation of classification pattern.

There are several methods to generate decision tree. ID3 is a well-known decision tree algorithm, which was proposed by Ross Quinlan of Sydney University in 1986. ID3 algorithm uses the entropy theory to select the attribute with the maximum information gain value in the current sample set as the test attribute. The algorithm of information gain value is: let s be the set of N data samples, and divide the sample set into m different test attributes [2].

Class CI (I = 1, 2,…, m), the number of samples in each class CI is Ni, then s is divided into m classes, and the information entropy or expected information is as follows:

Where Pi is the probability that the sample in S belongs to class I Ci. Suppose that the set of all different values of attribute a is values (a), and SV is the sample subset of attribute a in s whose value is v. at each branch point after selecting attribute a, the entropy of s classification for the sample set of the node is e (SV).The expected entropy caused by selection a is defined as the weighted sum of the entropy of each subset S, and the weight is the proportion of the samples belonging to SV to the original sample s:

$$E(S,A) = -\sum_{VEVvauws(A)} \frac{|S_V|}{|S|} E(S_V) \tag{1}$$

Where e (SV) is the information entropy that divides the samples in SV into m classes.

The information gain (s, a) of attribute a relative to sample set s is defined as:

$$E(S) = -\sum_{VEVvauws(A)} \frac{|S_V|}{|S|} E(S_V) \tag{2}$$

Gain (s, a) is the expected information compression of entropy caused by knowing the value of attribute The larger the gain (s, a), the more information the test attribute a can provide for classification. ID3 algorithm is to select the maximum information gain (s, a) attribute in each node as the test attribute.

3 Application of ID3 Algorithm

When freshmen enter school every year, colleges and universities will design some test questions based on the standard content of mental health test and their own specific conditions to investigate the mental health status of students. Schools with high degree of information have stored these data in the database of student management system or directly collected the data related to mental health through the network and stored them in the database, And the mental health of students in school also has tracking records stored in the database. Using ID3 algorithm in the decision tree, we can get the decision tree model from the "outdated" data in the database, that is, get the classification rules, and then use the classification rules to predict the future mental health status of the freshmen, so as to separate the freshmen who are most likely to suffer from mental diseases in the future, and list them as the key attention and assistance objects of the class teacher, student counselor, teacher, Put an end to all kinds of factors inducing mental illness, and let them grow up healthily in a harmonious and friendly campus environment. In this way, mental health education in Colleges and universities can be targeted and get twice the result with half the effort. The work of this paper is divided into the following steps.

$$E(S) = -\sum_{i=1}^{m} P_1 \log_2(P_1) + p(x_i) \tag{3}$$

3.1 Data Conversion and Cleaning

At present, the data in the existing database in Colleges and universities can not be directly used for data mining. Only by data cleaning and data conversion, can it be suitable for ID3 algorithm and improve the prediction accuracy of the model.

The purpose of data cleaning is to remove the noise and irrelevant information in the data set. For example, there are many fields in the mental health database. According to the prior knowledge, we can see that some fields (i.e. attributes) are not related to mental health, such as gender, age, native place and so on. Therefore, these fields can not be considered in data mining. It is helpful to build a better decision tree classification model by extracting the attributes that have a key impact on students' mental health, so as to achieve better classification and prediction effect [3].

The purpose of data conversion is to convert the data type and value of data source into a unified format. For example, the "family income" field in the mental health database was originally a continuous value. When data mining, this attribute must be converted to a discrete value (the attribute name is changed to "economic difficulty", and the value is yes or no) to be suitable for classification mining tasks. After data cleaning and transformation, the sample set that can be used to train the mining model (that is, to build the decision tree) contains 1000 samples, some of which are shown in Table 1.

3.2 Constructing Decision Tree With ID3 Algorithm

Prepare the sample set for training ID3 algorithm decision tree model.The decision tree is established by the following algorithm.

Table 1. Some samples in the training sample set.

Edit number	Attribute				Category
	Introversion	Family harmony	Hereditary diseases	Economic difficulties	Mental illness
1	No	Yes	Nothing	Yes	Nothing
2	No	No	Nothing	No	Nothing
3	Yes	Yes	Nothing	No	Nothing
4	Yes	No	Yes	No	Yes
5	No	Yes	Nothing	Yes	Nothing

1. For this sample set, the information gain of each attribute is calculated, because ID3 algorithm uses the information gain as the selection criterion of classification attribute. The larger the information gain, the more important the classification will be.
2. The attribute CI with the largest information gain is selected as the root node of the tree (or subtree).
3. The samples with the same value at CI are attributed to the same subset, and the value is taken as a branch of the tree. If there are several values of Ci, there are several subsets, and each value is taken as a branch of the tree.
4. Recursively call tree building for the sample subset with both positive and negative class examples [4].
   Algorithm.
5. If the subset contains only positive or negative examples, mark P or N on the corresponding branch and return the call function. After the completion of the tree building process, because the training sample set contains noise data, the decision tree generated is more complex.

3.3 To Predict the Mental Health of Freshmen

The decision tree classification model is used to predict the mental health of the freshmen, and the students with higher probability of suffering from mental health diseases in the future are screened out. The prediction results are distributed to the head teacher and student counselors in time. The screened students are regarded as the focus of mental health education, and timely and effective mental health counseling is carried out, As an important part of the construction of preventive college students' mental health education mode.

3.4 Application Integration

Application integration is to integrate ID3 algorithm (and other commonly used data mining algorithms) into the newly developed university management information system. For example, it is embedded in the university student management system based on

B/S structure, which is programmed with C# or VC++, and becomes a functional module of the system, The formation of students' mental health data collection, data processing, mining model construction, output mining report, new data sets for prediction and other functions of the perfect MIS system, and colleges and universities no longer need to buy expensive data mining software and hire professional data mining engineers for data mining work, so as to improve the information level of school student management.

The major of applied statistics is different from the major of statistics. Statistics is a methodology subject. It mainly studies the development and utilization of statistical information in theory, and cultivates students' professional knowledge of quantitative analysis and computer operation technology. Applied statistics is an interdisciplinary, comprehensive and application-oriented major. Its course content involves mathematics, statistics, economics and other fields. It has a wide range of applications in the fields of society, population, resources, commerce, finance, economy, pharmacy, epidemiology and engineering.

There are some problems in the teaching of Applied Statistics in Local Application-oriented Universities. On the one hand, it is mainly limited by its teaching resources, which directly affects the determination of teaching mode. For example, Shanghai University of Finance and economics, Ren min University of China and Zhejiang University of technology and industry are rich in teaching resources and teachers. They apply the curriculum system of statistics to strengthen the mathematical foundation, pay attention to statistical methods, and give consideration to the basic principles of economics. They are oriented to the application of government statistics, enterprise statistics, actuarial science, financial industry data analysis, macroeconomic and epidemic law exploration However, due to the lack of teaching resources in local universities, the practice teaching mode of professional practice teaching is relatively simple.

Only in the form of experimental courses and curriculum design. Through classroom practice, students can only master how to use specific statistical methods to solve certain problems under the premise of standard examples or known data. However, statistical investigation and analysis in reality are more complex. Therefore, when students face such comprehensive statistical investigation problems, it is difficult to find appropriate and appropriate solutions. Such teaching methods can not form a systematic teaching mode Practice teaching mode is not conducive to the cultivation of students; innovative application ability.

On the other hand, limited by the experimental materials, the effect of the experimental courses of applied statistics major in quite Local Application-oriented Universities is not ideal, mainly due to the failure to analyze the actual problems with the background of practical problems or according to the time-lapse materials or virtual data; for example, the statistical analysis of a stock is not based on the current economic environment and market trend Potential and other real situation of the stock for statistical analysis, but based on historical data on the trend of the stock for a simple analysis. In other words, due to the outdated teaching materials, it can not meet the requirements of the society for the experimental course of Applied Statistics. This will directly affect the students' application of statistical knowledge to solve practical problems, and then affect the quality of personnel training of Applied Statistics.

Based on the above ideas of teaching reform, the applied statistics major in Local Application-oriented Universities can carry out "project progressive" teaching construction. The teaching process design of "project-based teaching method" is essentially different from the traditional teaching method. The whole teaching process is no longer the link that teachers transfer the theoretical knowledge of textbooks to students through classroom teaching. Instead, teachers combine the theoretical knowledge of the course with practical problems based on the professional talent training objectives and Curriculum objectives, and design the teaching process according to the actual data Some feasible projects are planned, or combined with subject competition, these projects are decomposed into multi-step teaching tasks, and task-based teaching method is combined with project-based teaching method, so that students can complete the teaching tasks and then realize the operation of the whole project, and master the practical application of theoretical knowledge in this process.

The most important thing is that the role of teachers in the whole teaching process is no longer the dominant lecturer, but the guide and supervisor of students; learning process. In the early stage of teaching activities, teachers need to select teaching materials and decompose them into multi-step teaching projects; during teaching activities, teachers need to guide students to complete projects according to their project progress; in the late stage of teaching activities, teachers need to assess, summarize and evaluate the projects. In the process of completing the project independently, students can not only master all the teaching contents, but also cultivate the comprehensive skills of solving practical problems in cooperation with each other. In order to achieve the training of knowledge and ability as well as emotional goals, teachers carefully select teaching materials according to the training objectives of professional talents, curriculum syllabus, realistic social and economic phenomena and the northern boundary of subject competition, and decompose the selected teaching materials into multi-step teaching projects combined with teaching arrangement. Each project should contain the theoretical knowledge points of statistics course, It can also mobilize the enthusiasm of students to solve problems. The decomposed projects should be practical, operational and interesting.

Teachers can design teaching ideas according to the explanation of the theoretical knowledge used to complete the tasks in the project, that is, to complete the teaching in the form of "project". It is the basis and key to realize "project progressive" teaching to select the teaching project suitable for teaching and decompose the project into multiple teaching tasks. The following principles should be followed when selecting the project: first, the difficulty of the whole teaching project should be moderate. Teaching project design is too difficult or too easy will directly affect the enthusiasm of students to solve problems. Therefore, the selection of the project should ensure that it can be completed independently within the scope of students' learning ability. At the same time, it should also control the proportion of tasks that are less related to statistical knowledge or cannot be solved by students' current statistical knowledge in the project. Second, the content of the whole teaching project should be comprehensive.

4 Concluding Remarks

This paper introduces the application of decision tree in mental health education in Colleges and universities. It uses decision tree method to predict and classify the mental health status of Freshmen in the future, and finds out the most likely to suffer from mental diseases, so as to provide decision support for mental health education in Colleges and universities. The experimental results show that this method has a certain reference and application value for the construction of preventive mental health education mode for college students. The follow-up work of this topic is to find out more attributes that are highly related to students' mental health, and use other mining algorithms (such as neural network and genetic algorithm) to further improve the prediction accuracy.

References

1. Ren, Y., Zeng, S., Huo, R., et al.: Discussion and practice of new industrial Internet identification resolution system. Inf. Commun. Technol. Policy **8** (2019)
2. Li, B.: The core of intelligent manufacturing: industrial internet research. Mod. Inf. Technol. **2**(02), 191−193 (2018)
3. XCMG: XCMG information Xrea industrial Internet platform was selected as excellent product and application solution case of MIIT in 2017. Constr. Mach. **49**(6) (2018)
4. Shi, L.: Telecom operators' industrial internet strategy selection and strategy research. Inf. Commun. Technol. Policy (2018)