Design of a Comprehensive Student Information Management System Based on Data Mining Algorithms

Jiameng Zhang

{email: jiamengzhang112233@163.com}

Shandong Institute of Commerce and Technology, Jinan 250103, Shandong, China

Abstract. It uses data mining algorithms to classify and model a large number of things and then generate corresponding information for managers' decision making, so as to achieve efficient management of information and resource sharing. The system uses data mining algorithms to classify and transform students' problems in the learning process into corresponding knowledge according to different categories, and analyze them to get the optimal solution to improve school management. The system implements a comprehensive information management system for students, enabling administrators to perform statistics and analysis of all data. In this paper, the design of the comprehensive information management system for college students is based on data mining algorithms, and is analyzed and studied by analyzing the relationship between the functional modules and databases of the system, as well as the data exchange between the subsystems.

Keywords: Data mining algorithm. Integrated student information. Management system design

1 Introduction

With the rapid development of computer network technology, student information management system in colleges and universities has become an important part of student work. At present, many colleges and universities in China have adopted electronic attendance system, while electronic check-in, SMS and other functional modules are still in use, but it cannot match with the real working situation. Therefore, there is a need for an information management system that can automate daily business processing and meet the teaching needs of the school (referred to as "comprehensive student information management system"). A student information management system is a complete management system consisting of modules for student management, teaching, student registration, and comprehensive quality assessment. It needs to have the functions of collecting and entering different types of data (e.g. student registration files), matching the teaching content and teaching schedule with the school's teaching plan, so that it can realize the correlation analysis between various data and form personalized education programs. As most universities are currently operating in the "manual+automation" mode, the whole system requires a large and complicated manual operation. In order to solve this problem, this paper proposes a data mining algorithm-based comprehensive student information management system construction plan.

ITEI 2022, December 23-25, Harbin, People's Republic of China

DOI 10.4108/eai.23-12-2022.2329191

2 Research on Data mining Technology

2.1 Data mining Technology

Data mining technology refers to the extraction of valuable information from massive, fuzzy databases and the classification of these useful resources through association rules, thus enabling the analysis of potential user needs and behavioral patterns as well as other implied relationships or processes. Different fields have very different approaches to the study of related problems, so there are different requirements for data processing techniques [1]. At present, there are two main approaches: one is to discover potential objects of interest from the massive original sample database, then use data mining techniques to reorganize and filter the information in the original sample database, and finally identify potential users or similar objects that have been found and exist, and finally classify them. The other one is to mine variables with the same attributes from multi-dimensional variable space classification, thus making a new model and making a new data warehouse. This method can transform between different models, but the process is determined by the amount of information and the speed of processing [2].

2.2 Data Mining Methods

2.2.1 Decision Trees

Decision tree can be regarded as a tree classification model, which classifies objects by using a tree structure and classifies them by dividing them into decision trees, thus realizing the analysis of the relationship between data objects and the corresponding information variables in the database. The decision tree uses a top-down recursive approach to recursively classify data objects so as to predict the information variables in the database as a whole, and determine whether the decision tree is the optimal class based on the results. The comprehensive evaluation of the student learning process and the system performance indicators and running time conditions reveals that the performance indicators and running time of the integrated information management system can meet the user requirements for system development, and it can also guarantee the time required for the data processing process.

Decision trees are derived from the concept learning system CLS, and the ID3 algorithm was first proposed, and then the C4.5 algorithm, which can handle continuous attributes, emerged after improvement, and then this algorithm can be used to deal with discrete classification problems.

(1) ID3 algorithm

ID3 algorithm is the most classical and influential algorithm in decision tree algorithm, which can divide from a tree table into many children in a continuous space, and then classify each subset to get the optimal solution.ID3 algorithm introduces the concept of information first in information theory into decision tree algorithm, and realizes the classification and reorganization of information, and takes it as a whole to facilitate the decision maker to make the corresponding choice. When building the internal nodes of the decision tree, the attribute with the highest information gain value in the training set is selected as the test attribute, and it is used as the output variable. Finally, based on the decision tree, the data mining algorithm is improved accordingly, and auxiliary functions such as data pre-processing module and text clustering analysis module are designed so that system information management, database

maintenance and statistical query operations can be realized, and finally a decision tree model that can classify the unclassified data set is obtained [3].

Firstly, the definition of information first and information gain are given, and the information gain is defined, and then a mapping relationship is established between the class A dataset and the class B document collection according to the characteristics of the algorithm. Let there be s data samples in the dataset S. Assume that the category attributes have m different values, and let the number of samples with category Ci in the dataset be si. The expected information required to classify the dataset is called information first, as shown in Equation (1):

$$I(s_1, s_2 \cdots s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Let attribute A have v distinct values {a1,a2,.... av,}. Use attribute A to divide data set S into v subsets {S1,S2... Sv}, where subset Sj represents all samples whose attribute A values aj in data set S. According to the relationship between the sets in the data set, the sorted subset is obtained, and then they are classified, and all the classes are divided into all the objects contained in this category. Finally, an example is given to illustrate that this method can effectively improve the efficiency and quality of decision making and provide better service for users [4]. Calculate the information entropy of attribute A divided into the corresponding subset as shown in Formula (2) :

$$E(A) = \sum_{i=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j} \cdots s_{mj}) \tag{2}$$

As can be seen from the formula for calculating information entropy, the calculation of expected information of the corresponding subset si is shown in Formula (3) :

$$I(s_{1j}, s_{2j} \cdots s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \tag{3}$$

The information gain value of the data set S divided by attribute A can be calculated and compared with the information gain value of the original data set to obtain the optimal solution, as shown in Formula (4) :

$$Gain(S, A) = I(s_1, s_2 \cdots s_m) - E(A) \tag{4}$$

In short, ID3 algorithm realizes the classification, mining and classification of data by calculating the information gain value of each attribute in the data set, and can carry out information management according to different types of users, providing a new mode for the comprehensive information management system of school students [5].

(2) C4.5 algorithm

C4.5 algorithm is an optimization algorithm based on ID3 algorithm proposed by Quilan in 1993. It is an algorithm based on computing technology. The algorithm mainly solves the problems of incomplete information, data redundancy and non-real-time system, but it has defects for the high-dimensional, inefficient and unstructured model. That is, there may be a large number of invalid information or errors in the process of processing, and due to the large time complexity

and space dimension, it cannot effectively conduct classification management and output analysis. Therefore, the algorithm needs to be improved to deal with high dimensional data sets and unstructured information flow better. C4.5 algorithm is shown in Figure 1:

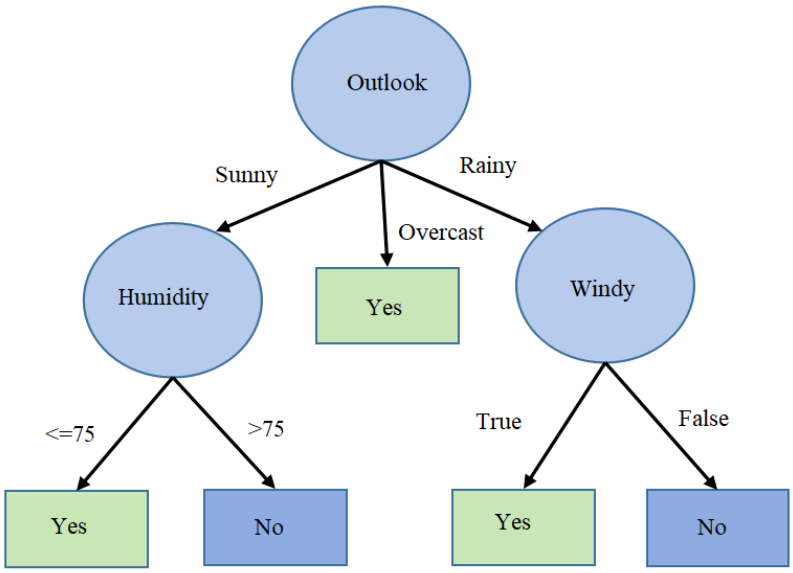

Figure 1. C4.5 algorithm

When establishing nodes, C4.5 algorithm selects the one with high information gain rate from the attributes of the data set as the test attribute, and takes the value of the attribute set as the output variable. In the data set, the algorithm generates the corresponding comprehensive information, so as to improve the decision-making efficiency of the system. Suppose attribute A has n different values {a1,a2,.... ,an}, and use attribute A to divide data set S into S1,S2,.... ,Sn has n subsets, then the information gain rate of sample set S divided by attribute A is calculated as shown in formula (5) and (6) :

$$GainRatio(S,A) = \frac{Gain(S,A)}{Split(S,A)} \tag{5}$$

$$Split(S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2\left(\frac{S_i}{S}\right) \tag{6}$$

(3) CART algorithm

CART algorithm uses binary recursive partitioning principle, that is, the current data set is divided into two subsets, each subset contains all the calculated data, and each set is represented by a candidate data set, wherein the number of variables in the subset is compared with the current input information. The candidate data sets are divided into different subsets and the next control policy is decided based on the current state [6].

Suppose for the data set T, $gini(T) = 1 - \sum p_j^2$, where pj represents the probability of occurrence of category j in T. If T is divided into T1 and T2 subsets, gini coefficient can be calculated, as shown in Formula (7) :

$$gini_{split}(T) = \frac{S_1}{S} gini_{split}(T_1) + \frac{S_2}{S} gini_{split}(T_2) \tag{7}$$

ID3 algorithm and C4.5 algorithm only set category attributes of leaf nodes, while CART algorithm sorts according to specific information characteristics in the system and takes them as basic attributes, and then synthesizes each ID3 algorithm and CART algorithm.

2.2.2 Neural network

Neural network is a new algorithm formed by a large number of neurons through the process of information processing in the biological nervous system. It has a high degree of parallelism and adaptive ability, and it can effectively solve complex problems. Artificial neural network mainly includes the following steps: learning, recognition and selection. First of all, it is necessary to determine the connection between the input layer and the output sample units, determine the type and quantity of input sample units, and then learn from them, and divide the network into different areas according to the output results. Secondly, weight assignment is performed on each training set and its average value is calculated as the output result, which is output to the neural network as the input result. Finally, the weight coefficient is adjusted according to the actual data to meet the requirements of the system and realize the dynamic optimization of the neural network [7]. The neural network algorithm is shown in Figure 2:

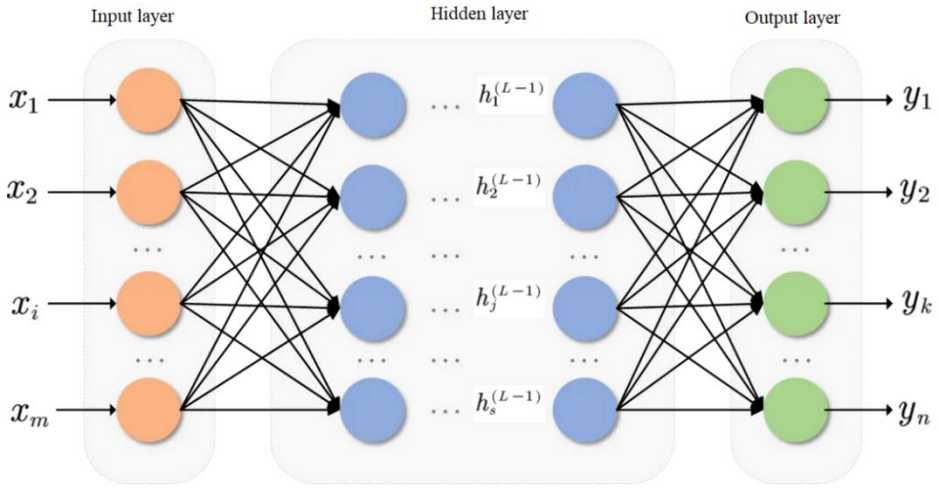

Figure 2. Neural network algorithm

2.2.3 Genetic algorithm

Genetic algorithm is a random search method proposed based on the theory of natural evolution. Its basic idea is as follows: First, an optimal solution is selected from the three links of selection, crossover and mutation as the initial population, and then the optimal individual is copied

according to certain rules to get the required objects in the next stage. In the practical application process, it is found that many kinds of uncertain information are generated after the generation of a large number of data sets, which often have obvious defects or strong randomness and cannot be predicted. Therefore, genetic algorithm has the advantages of high search efficiency, short time and good robustness, but it also has some limitations. For example, high dependence on data set, unable to adapt to complex environment and multi-source information [8].

3 Design of Comprehensive Information Management System for Students

3.1 System Architecture Model

B/S structure is based on data centralization and processing as the core, in the design process it adopts the idea based on object oriented, and it is built as an infrastructure. The system mainly includes database management, user access control, security and other modules, which are implemented by the background program. After the completion of the front-end application program, it executes commands to the client and accepts corresponding requests, while the background is responsible for receiving the required data information and feedback results returned from the server to the development environment. Then according to the actual needs of the software to modify or add functions and delete processing, so as to meet customer needs. The background program is executed by the operator in the development environment, and the administrator needs to maintain the information of the user in the management process [9].

3.2 Overall Functional Structure Design of the System

All the functions of this system are completed through the interaction between the application server and the user browser. The student integrated information management system based on data mining technology is a very complex and valuable system. The overall structure of the system is shown in Figure 3.

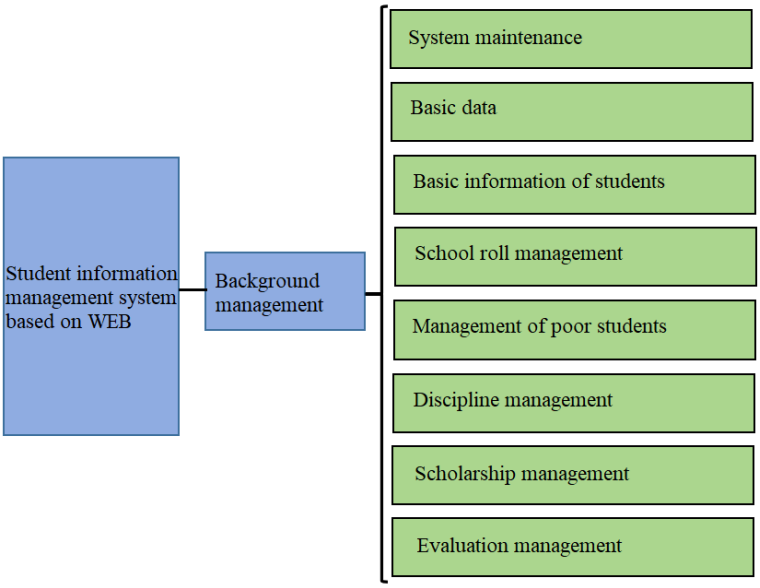

Figure 3. Overall structure of the system

4 System Test

4.1 Test Environment

Software testing is a testing process that provides reliability and integrity for developers. Errors are found and corrected in time through a series of methods, so as to improve system quality. The test environment of this system is shown in Table 1:

Table 1. Test environment

	Name	Hardware or software
System development hardware environment	Processor (CPU)	Celeron III + processor
	Computer memory	More than 1G
	display	VGA compatible display system (1024*768)
	Hard disk capacity	Hard drives above 20 GB and 7200 RPM /s
	Optical drive	DVD with burn function
	printer	A laser printer of any model
System development software environment	Operating system	MS Windows XP or Windows7
	Database management system	The platform can adapt to Microsoft SQLSERVER and Oracle
	Development platform	Microsoft Visual Studio 2008

4.2 Test Method

Test method is a very important part of the program design, including white box and black box, white box test is a method in the program design, it is mainly based on the software code packaging test out the internal error. Black box test is a kind of test based on system function. It mainly checks the function of the program code by detecting whether the data exchange and input and output between the modules of the system are correct [10]. The white box and black box test representations are shown in Figure 4:

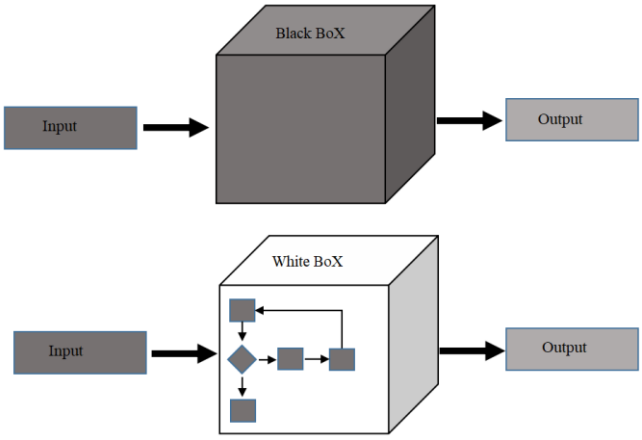

Figure 4. White box and black box tests

4.3. Test Cases

The purpose of test cases is to ensure the normal operation of the system, analyze and design it, and finally meet user requirements. In this paper, the user login test case is taken as an example, as shown in Table 2:

Table 2. User login test cases

Use case number	Test characteristic	Use case description	System reaction	Test conclusion
YHDL_001	Functional test	Use User name: admin; Log in to the system with password 1234 and check the login status	Login success	pass
YHDL_002	Functional test	Log in to the system using a user name other than admin and password 1234 to check the login status.	Login failure	pass
YHDL_003	Functional test	Log in to the system as user name: admin: Password 1234 and check the login status	"Password login Failed" is displayed.	pass
YHDL_004	Functional test	Log in to the system using a user name other than admin and a password other than 1234 to verify the login status.	Login failure	pass

5 Conclusion

After testing and analyzing the system, a comprehensive information management system based on data mining algorithm is proposed in this paper, and the software is developed. First of all, this paper writes a detailed function module according to the feasibility analysis, demand demonstration and design stage, and then through the design to achieve a complete and meet the requirements of the user's comprehensive information system, and completed the development of the software. The test results show that the proposed method is feasible, effective, stable and reliable, and has good scalability.

References

[1] Papi Ramin,Attarchi Sara,Darvishi Boloorani Ali,et al. Knowledge discovery of Middle East dust sources using Apriori spatial data mining algorithm[J]. Ecological Informatics,2022(1):70-72.

[2] Wang FeiXiang,Ji Rui,Zhang LuMing,et al. Pelvic Injury Discriminative Model Based on Data Mining Algorithm.[J]. Fa yi xue za zhi,2022(3):38-40.

[3] Niu Wenjia,Zhao Lihua,Jia Peiyao,et al. An Audit Risk Model Based on Improved BP Neural Network Data Mining Algorithm[J]. Advances in Multimedia,2022:20-27.

[4] Anonymous. Campus Management Corporation; Campus Management and Ad Astra Partner to Deliver Real-time, Two-way Integrated Student Information System and Scheduling Solution[J]. Computer Weekly News,2010:23-28.

[5] Lempert Jeremy,Kollmeyer Phillip J.,He Melissa,et al. Cell selection and thermal management system design for a 5C-rate ultrafast charging battery module[J]. Journal of Power Sources,2022:550.

[6] Ma Dehou. On-the-Spot Decision-Making System of Basketball Game Based on Data Mining Algorithm[J]. Security and Communication Networks,2022:37-39.

[7] Laayati Oussama,El Hadraoui Hicham,Guennoui Nasr,et al. Smart Energy Management System: Design of a Smart Grid Test Bench for Educational Purposes[J]. Energies,2022(7):12-15.

[8] Tarhan Burak,Yetik Ozge,Karakoc Tahir Hikmet. Hybrid battery management system design for electric aircraft[J]. Energy,2021:234.

[9] Junkui (Allen) Huang,Shervin Shoai Naini,Richard Miller,et al. Unmanned autonomous ground hybrid vehicle thermal management system: design and control[J]. International Journal of Vehicle Performance,2020(3):14-16.

[10] Wang FeiXiang,Ji Rui,Zhang LuMing,et al. Pelvic Injury Discriminative Model Based on Data Mining Algorithm.[J]. Fa yi xue za zhi,2022(3):38-42.