Development and Research on the Analysis System of Physical Quality of College Students Based on Data Mining and Decision Tree Algorithms

Xuelei Zhang\textsuperscript{1}, Xiangsheng Meng\textsuperscript{2}, Yao Ma\textsuperscript{1}, Changwei Wang\textsuperscript{3}

\textsuperscript{1}Institute of Disaster Prevention, Sanhe 065201, China
\textsuperscript{2}Yanching Institute of Technology, Sanhe 065201, China
\textsuperscript{3}Ningxia University SportS Institute, Yinchuan 750021, China

Abstract

Data mining refers to find out data information related to research theme from a mass of data, is an all-around and in-depth data analysis mode, and has been widely used in lots of fields at present. The decision tree algorithm is a process of comparing and analyzing data in details through a series of calculation rules, and finally classifying and completing the detailed analysis and integration for data warehouse and data mining technology of the research object, and then getting systematic and scientific research conclusions. This thesis will analyze and study on the development of the analysis system of college students’ physical quality based on data mining and decision tree algorithms.

Keywords: Data Mining, Decision Tree Algorithm, Physical Quality of College Students, Analysis and Research.

1. INTRODUCTION

Thanks to the continuous development of information technology in our country, the teaching level of physical education of the higher education institutions also get improved comprehensively, and meanwhile, the colleges and universities begin to attach importance to the quality development of physical education of college students, some of them have had comprehensive records and statistics for PE achievements and physical quality data of colleges students. In face of a mass of data of students, the traditional database management mode and data statistics means cannot meet the requirements of the physical education of colleges and universities. But the data mining and data warehouse technology can extract the concealed information that unknown previously and own potential value from a mass of students’ physical education data with high dimension and uneven quality, and then assist colleges and universities to integrate and manage the information of physical quality of college students (Xie, 2013).

Meanwhile, the data mining in physical education of college students can be applied to find and solve the problems timely, can analyze data in depth, provide the information required timely and accurately, and also help us to find out the interrelation of various elements in a great number of data, and find some new rules of guiding the scientific research, teaching and classroom training, which are generated with the dynamic changes of many factors. Therefore, during specific practical teaching management process, how to use data warehouse and data mining technology to realize informatization management and depth analysis on the data information of physical quality of students is an important subject that the scientific researchers of physical education need to study, and also the inevitable requirement that college physical education adapts to the science and technology development of modern information better (Li et al., 2013).

2. RESEARCH OVERVIEW

2.1 Background overview

With the continuous development and progress of information technology and scientific and technological level in our country, and the implementation of a series of measures of deepening the reform of the college education...
system, the quality of physical education of colleges and universities of our country has been improved comprehensively. The practical teaching of public physical education of colleges and universities has been reformed and developed from the aspects of concept, contents and teaching mode and methods. At present, the study contents related to the practical teaching of public physical education of colleges and universities mainly involve the following two aspects: the studies on the reform and innovation of the practical teaching methods of public physical education of colleges and universities, and the evaluation and improvement for the physical quality and health of college students. Colleges and universities establish relatively comprehensive data statistics record for PE achievements and physical quality, in face of a mass of students’ physical education and health data, the original database can't meet the demand of management mode and data statistics, so that colleges and universities fail to build transversal connection relation among various physical education data during the process of physical education, cannot respond to the national development strategy of health physical education and the requirements of offering supports for the competition of sports competition talents (Liu, 2016). Therefore, colleges need to combine data mining and data warehouse with specific research and practice while analyzing the physical quality of college students, which can not only carry out more systematic and scientific statistical analysis on the data, but also assist colleges to accurately find the attributes of the problems in physical education and health of college students during the process of data analysis, and also analyze data in depth.

2.2 Literature review

Data mining, also known as data exploration, is one of the programs in Knowledge-Discovery in Databases. It usually refers to the process of searching out the information concealed in a great number of data through calculation, such process needs to be completed through connection to computer programs, and requires data statistics, on line comparison, intelligence search, expert operation and cognitive pattern and many other approaches to complete the targeting of data (Li and Wang, 2016). The development of data mining technology originates from the thought of many fields, such as, sampling judgment, estimate judgment and hypothesis testing in statistics; the search calculation, the idea of mathematical modeling and scientific theories in artificial intelligence recognition system and intelligent learning. As data mining is rapidly integrated with the thoughts from other fields, which include the most efficient evolutionary computation, information theory, signal processing, visualization and information resource retrieval, the requirements for database becomes indispensable, the database is able to provide safe storage space, convenient indexes and query processing support for data mining. The high-performance technology of calculation will play an irreplaceable role in processing a mass of dataset; the distributed technology can also help to process mass data in a centralized way, which is one of the most important part (Xie et al., 2015).

The decision tree algorithm is a method close to the calculation of discrete functions, and owns characteristics of typical classification, at first, the decision tree algorithm will utilize induction algorithm to generate readable rules and decision tree model through data processing, and then use decision-making computation formula to analyze new data systematically. In essence, the decision tree algorithm is the process of carrying out the comparative analysis on data in details through a series of calculation rules, and finally completing data calculation for research objects (Chen et al., 2012). In the construction of the decision tree algorithm, the most efficient data classification rules will be usually selected to build small-scale decision tree with high precision. The construction of the decision tree algorithm can be divided into two steps, firstly, the generation of a decision tree is the process of forming decision tree out of sample dataset, and the sample dataset is formed according to practical demand or the combination of historical data, and then processed through data analysis technology. Secondly, it is the pruning process, the pruning of the decision tree is the process of testing, correcting and modifying the generation process of upper stage, mainly refers to verify the initial shape generated during the process of the generation of decision tree through sample data and data concentration, and prune these branches that will affect the accuracy of prejudgment.

3. DEVELOPMENT AND RESEARCH ON THE ANALYSIS SYSTEM OF THE PHYSICAL QUALITY OF COLLEGE STUDENTS BASED ON DATA MINING

3.1 Research process for data mining of the physical quality of college students

The data mining process can be mainly divided into the following seven steps:

(1) Data cleaning: clean out the data obviously irrelevant to the mining theme of the physical quality of college students.
(2) Data integration: combine with multivariate data, and form the data warehouse of physical quality of college students.

(3) Data selection: extract and select data related to the physical quality of college students from data warehouse.

(4) Data conversion: convert data to be the data format easy for data mining.

(5) Data mining: utilize the data mining method to carry out the depth analysis on data rules or knowledge.

(6) Evaluation model: screen out related knowledge of the model of physical quality of college students from mining results according to evaluation standard.

(7) Expression knowledge: utilize visualization and knowledge representation technology to integrate the data of the physical quality of college students obtained through data mining.

Figure 1. Data Mining Technology

3.2 Application of data mining in the analysis system of physical quality of college students

It can be observed from the definitions of data warehouse and data mining that, it is the expansion and extension of statistical analysis methodology, utilize data warehouse and data mining technology to mine the above data (Yin et al., 2016). The data mining technology can be applied in the field of college physical education flexibly, such as, teaching management of physical education, teaching evaluation of physical education, curriculum provision of physical education, teaching methods of physical education, and selection of teaching material. The data resource technology on data mining can be utilized sufficiently to build the evaluation model of physical quality according to different features of students, such as, gender, age and physical function and other data, and thus find the reasons for the problems of physical quality, and then improve the teaching methods of college physical education and increase the quality and effect of physical education. Therefore, the researchers in the field of college physical education can put forward the definition of the problems according to the accumulated data currently, utilize mining tool to build model for solving problems, and then evaluate, verify and implement the model.

3.3 Research conclusions

Through comparative analysis on data, the researchers in the field of college physical education utilize mining tool to build model for the solution of the problems, and then evaluate, verify and implement the model and come to a conclusion (Yang et al., 2016). Based on data warehouse and data mining, the results of the mining of mass data in the field of college physical education show that, the physical quality of college students in our country is quite lower at the present stage, the students are lack of interest in physical exercise, and thus most of them stay in the sub-healthy state. On the basis of data mining and data bank and basic procedures of data mining, the thesis analyzes and studies the physical quality of college students, come to a conclusion that colleges and universities shall pay attention to the physical education and daily amount of exercise of college students, and record the data of their physical fitness and physical education curriculum every year.

4. DEVELOPMENT AND RESEARCH ON THE ANALYSIS SYSTEM OF THE PHYSICAL QUALITY OF COLLEGE STUDENTS BASED ON DECISION TREE ALGORITHM

The decision tree algorithm was proposed by Hunter firstly; with the continuous development of big data technology, C4.5 has been applied into the classification algorithm of big dataset, i.e. C5.0 algorithm. Such kind
of algorithm becomes more optimized on the aspects of execution efficiency and internal storage, the data calculation of the development and research on the analysis system of the physical quality of college students in this thesis is done by adopting C5.0 algorithm of the decision tree.

4.1 Data processing

The decision tree algorithm C5.0 can be used to process a great number of data, such as, Date and Times, such kind of algorithm shows higher performance on the processing speed and classification precision, and does not need to estimate after many times of training, and the data deduced from the model is very intuitive, but C5.0 usually has higher requirements on successional fields, therefore, we need to select the information gain with higher attribute to determine the standard of the branches of decision tree in the development and research of the analysis system for the physical quality of college students, in order to seek for the optimized grouping variables and the fixed points of segmentation. It is assumed that S is the set of samples, and has K classifications in the object variable C, freq (C, S) means the sample data belongs to the classification of Ci, in which |S| represents the sample data of the sample set S, and the information of the set S is defined as:

\[ \text{info} (S) = \sum_{i=1}^{k} (\frac{\text{freq} (C_i, S)}{|S|}) \log (\frac{\text{freq} (C_i, S)}{|S|}) \]  

(1)

In which, if a certain attribute variable T has N classification sets, after the attribute variable T is introduced, the definition changed to be:

\[ \text{info} (T) = \sum_{i=1}^{n} (\frac{|T_i|}{|T|}) \times \text{info}(T_i) \]  

(2)

The information gain value of the attribute variable T becomes:

\[ C_{ain}(T) = \text{info}(S) - \text{info}(T) \]  

(3)

It is assumed that tree () is the generation function of the decision tree, set R to be nonstandard attribute set, C to be standard attribute set, and S the training set:

The function tree(R, C, S)// decision tree is the returned value of the function, in which:

* \{d_j\} \quad j = 1, 2, ..., m 
* \{S_j\} \quad j = 1, 2, ..., m 

is the value of the attributes D and S,

In which, If(S is null) then it is failure to return the next node; If (the attribute value of record marks included in S is equal) then it returns to the next node of the attribute value of such mark; If(R is null) then the single node of the equilibrium value of S is used when it returns;

/*The data value is error at this moment, data cannot be classified normally*/

Find the information gain value D with the maximum attribute, and generate a big tree with the root of D, the branches are respectively d_1, d_2, ..., d_m; utilize the recursive call function tree \((R - \{D\}, C, S_1)\); tree\((R - \{D\}, C, S_m)\) to realize the data construction of decision tree for the development and research of the analysis system of the physical quality of college students (Lu and Zhao, 2014).

4.2 Data preprocessing

When C5.0 is applied in the development of the analysis system for the physical quality of college students, it needs to preprocess data. The source of data comes from the score table of physical education test for 5065 students Grade 2105 of a college in Heilongjiang Province in the school year 2015-16, including 2802 boy students and 2263 girl students; the physical quality of college students can be judged through score analysis on the scores of physical education tests of college students (Huang, 2017).

The data of the physical health of college students in such college have been investigated and sorted out, please see details in Table 1, through combination and comparison with the scores of physical education tests of college students.
students, the more valuable and meaningful data can be obtained. And then these data are processed, the data cleaning mainly refers to process the missing value of data, the abnormal data, repeated data and noisy data, and remain the effective data for model building, for example, the data cleaning for the names and contact information of parents in basic information can be conducted (Zhou, 2017).

### Table 1 Physical Health of College Students

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Literature and history</th>
<th>Science and engineer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>1006</td>
<td>950</td>
<td>1108</td>
<td>1029</td>
</tr>
<tr>
<td>Sub-health</td>
<td>1456</td>
<td>1000</td>
<td>1452</td>
<td>1002</td>
</tr>
<tr>
<td>Unhealthy</td>
<td>340</td>
<td>313</td>
<td>246</td>
<td>232</td>
</tr>
</tbody>
</table>

#### 4.3 Construction of the decision tree model in the analysis system of the physical quality of college students

The relatively authoritative data mining software includes SAS Enterprise Miner, SPSS Clementine, Oracle DM, IBM Intelligent Miner and SQL Server Data Mining, etc. When the decision tree model for this analysis system of the physical quality of college students is built, we select SAS Enterprise Miner 2.0 as the platform for model building and analysis; there are 4 kinds of calculation methods of the decision tree: C5.0, CART, QUEST and CHAID, in which QUEST and CHAID are taken as the branches of decision tree, and C5.0 and CART as multiple branches of decision tree (Wei and Tian, 2014); the algorithms of CART and QUEST could be the continuous object variables, or discrete C5.0 or QUEST algorithms, therefore, the data streams through the data mining of the decision tree in the analysis system of the physical quality of college students are as follows:

![Decision Tree C5.0 Mining Data Streams](image)

#### 4.4 Research conclusions

Through the building of the decision tree C5.0, the QUEST algorithm is used to complete data calculation of the decision tree and deeply mine the data of the physical quality of college students, and analyze the solutions for the physical fitness of college students according to the results of data mining. Through the development and research on the analysis system of the physical quality of college students, we can lean the physical fitness condition of college students at present; therefore, the colleges and universities shall attach importance to train the physical quality of colleges, solve the problems in the physical education of college students specifically, and improve the specialized team and mechanism for the physical training and assessment of college students, and meanwhile, the colleges and universities shall make good use of network, broadcasting, campus newspaper and other methods to publicize physical education, and carry out various forms of publicity activities of physical education. And it also needs to conduct the trainings on the professional skill of the teaching staffs of physical education in colleges and universities, and reinforce the construction of teaching staffs of physical education in colleges and universities. The colleges and universities need to pay attention to the communication with the parents of students, and let parents sufficiently realize the importance of training physical quality (Liu, 2012), and also positively guide the development of college students on the aspect of physical education.

#### 5. CONCLUSION

The above development and research on the analysis system of the physical quality of college students based on data mining and decision tree algorithm is only a simple experiment for the application of decision tree algorithm and data mining technology in the field of physical education in colleges and universities. During the development process of college education reform in our country, these two kinds of technology will be applied in the teaching data analysis more frequently and generally, and will also bring fresh vitality to the reform and development of the cause of physical education in colleges and universities. The development and research on the analysis system
of the physical quality of college students through data mining have a certain guidance meaning and reference value for the researchers of physical education discipline and data mining researchers in our country, such highly efficient, accurate and timely data supply can help us to find the required and correlative factors in the gigantic database, and understand the problems that exist in colleges according to the changes of these dynamic factors, and then adjust and change the teaching model and concept of physical education for college students (Huang, 2017). Therefore, during the practical process, how to utilize data warehouse and data mining technology to achieve the informatization management and depth analysis on the data of the physical quality of students, realize the perfect combination of modern technology and the cause of physical education and promote the improvement of the physical quality of college students are the important contents of the development of the educational cause at the present stage.

ACKNOWLEDGEMENTS

Experimental research project of higher education teaching reform in Hebei province: The pathological diagnosis and correction of the reform of PE curriculum in colleges and universities (2016GJJG250).

REFERENCES

Chen S.P., Yan Z.L., Han J.L., Zhu C.Y. (2012). College students physical health standard data mining - grade features and exercise behavior among the students of Xi’an Jiao Tong University as an example, Sports research and education, 27(01), 122-125.
Li W.P., Qi D.Q., Cai J, Wei H., Lei W. (2013). The sports consumption structure and characteristics of urban residents in Xi’an City, Data mining perspective sports science, 33(09), 22-28.
Xie X.Y. (2013). Research and application of data mining in the analysis of sports data, Contemporary sports science and technology, 3(23), 9-10.
Yin D.F., Zhong Y.P., Yu Y.G. (2016). In the application of data mining technology in human muscle power data analysis on the testing data of human grip muscle strength as an example, Sports science, 30(02), 70-74, 82.