Physical Analysis and Research Based on Data Mining

[1,a] Zhu Xueqiang

[1] Shandong Sport University , Jinan, ShanDong 250102,China

[a]zxqzmx@163.com

Keywords: Data mining, physical training, model, decision tree.

Abstract. In order to enhance effectiveness and correctness of physical training, the paper, based on data mining, puts forward data processing flow and overall framework model in physical training process, and conducts statistics and analysis of ID3 decision tree algorithm on physical training's plenty of data information. The experiment indicates that the algorithm can better conduct data analysis on physical training and provide training assistance for teachers.

Introduction

In recent years, data mining has already become a research hotspot in computer realm. With the extensive application of information technology in physical exercise, a large number of selection materials concerning growth of athletes are gathered gradually [1, 2, 3]. Real and valid data on training and competition can obtain hidden knowledge and rules by settling and analyzing data, so as to provide useful assistance for serving researches on physical education, acquiring better performance of athletes, decision and management of physical administrative departments.

Sports science possesses lots of data information resources and sports statistics concerning training, teaching and competition. How to utilize these valid data and discover potential and available principles is one of the problems for sports science that is urgent to be solved by using computer technology [4, 5]. Existing data mining may provide assistance for this purpose. It refers to a process that excavates hidden, unknown and useful for decision knowledge from large-scale data concentration. By taking advantage of data mining, it can promote its training and service level in physical education, physical training and sports competition and can better satisfy requirements of various sports science researchers at different levels. However, essential data can't be filled in perfectly or accurately, the data itself possesses uncertainty, modeling scale of data warehouse is too complicated, and there is design flaw in mining algorithm. The above-mentioned reasons cause distortion of excavation results, so coaches and managers can't be convinced completely. Therefore, it can't play a guiding role on every practical link.

In order to excavate core competitiveness of college physical education for optimizing resource allocation of college physical education, and improving resource utilization rate, administrators in college physical education should master the development direction of college physical education in rapid and variable competition, and analyze data mining and management information-based current situation of college physical education. The paper explores the application of software technologies in athletic performance management, for the purpose of enhancing efficiency of athletic performance. Moreover, the paper analyzes athletic performance by using data mining, serving for the improvement of sports teaching quality.

Data Mining

The increasingly increased data hides much important information. People hope to analyze it on a higher level, so as to take advantage of these data better. In order to provide a unified overall situation for decision makers, the data warehouse is established in many fields. However, lots of data often make people can't distinguish the information that hides in it and can provide support for decision, while traditional inquiry and reporting tools can't satisfy demands of excavating these information. Therefore, it needs a kind of new data analysis technique to deal with lots of data and extract valuable potential knowledge from it. Consequently, data mining emerges at the right moment from this. Data mining is also gradually perfect with the development of data warehouse. The data mining function is to discover potential rules and knowledge from databases to enrich contents of knowledge base, offer more decisions and ensure accuracy and practicability of decisions.

Researches of data mining combines with technologies and achievements of multiple different subject areas, making current data mining methods present diverse forms. From the perspective of statistic analysis, linear analysis and non-linear analysis of data mining model, regression analysis, logistic regression analysis, univariate analysis, multivariate analysis, time series analysis, nearest sequence analysis, nearest neighbor algorithm and cluster analysis, etc. methods are utilized in statistic analysis techniques, which can examine those data with abnormal forms. Next, taking various statistic models and mathematical models to explain these data, and explain market disciplines and business opportunities hiding behind these data. Data mining of knowledge discovery is completely different from the data mining of statistic analysis, including artificial neural network, support vector machine, decision tree, genetic algorithm, rough set, rule discovery and relevant sequence, etc. This system adopts the data mining of knowledge discovery, analyzes data in line with corresponding model of users' input selection, enriches knowledge base and model base in accordance with new knowledge and model generated by data mining, produces rational solution of exercise training gradually, according to existing knowledge base and model base, and returns to user interface ultimately.

Data Mining Process and Framework Model of Physical Training's Data

Data Mining Process: Data mining is a process that makes use of data mining technology to excavate valuable knowledge from large-scale data of database, data warehouse and other information database. The integrated data mining process contains data standardization, data integration, data transformation, data mining, pattern evaluation and knowledge representation, etc. several steps, which can be divided into four layers, as shown in Figure 1. The data in college sports education sets up data warehouse and directory by using data management process of standardization, cleaning and noise reduction and executing designing scheme of data warehouse, translates it into data mining base, data mart or data mall, becomes excavation objects, discovers knowledge or pattern by utilizing data mining tools and methods, such as artificial neural network, genetic algorithm, decision tree, nearest algorithm and rule deduction, etc, and sends results to terminal users in college sports management after explanation.

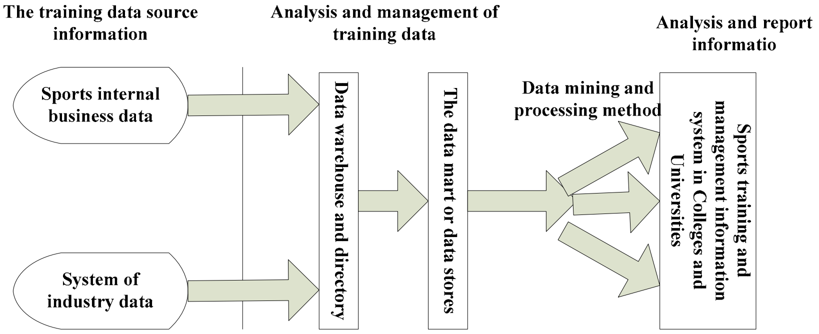
The training data source information
Analysis and management of training data
Analysis and report informatio
Sports internal business data
System of industry data
Data warehouse and directory
The data mart or data stores
Data mining and processing method
Sports training and management information system in Colleges and Universities

Figure 1 Data Mining Process Diagram of Physical Training

Overall Framework Model of Physical Training's Data Mining System: First of all, facing to day-to-day business, it is necessary to construct integrated online transaction processing database system to provide data for data mining, form operational data for data source, give priority to build management information system, acquire data from day-to-day business work, embody specific process of sports work, realize database storage of business data, and provide conveniences for data maintenance, adding, deleting, modifying and inquiring, such as office automation, teachers management system, scientific research management system, sports teaching management system, curriculum arrangement, curriculum selection, question bank, performance, teaching evaluation, athletic competition management system, asset management system, stadium, equipment, financial management system, physical test management system, social service management system, sports team management system, group campaign management system and intelligence system, etc.

Secondly, facing to business problem, it needs to build online analytical process data warehouse to form analytical data, mainly give priority to construct decision support system, acquire data from data center, establish data warehouse in line with decision-making theme, and apply CRIPS process model and cross industrial data mining process to solve problems of administrative decisions, as shown in Figure 2., which is overall framework model based on data mining system.

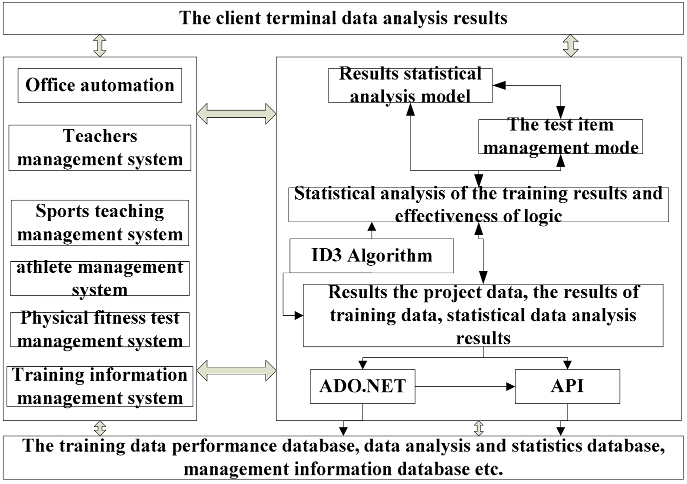
The client terminal data analysis results
Office automation
Teachers management system
Sports teaching management system
athlete management system
Physical fitness test management system
Training information management system
Results statistical analysis model
The test item management mode
Statistical analysis of the training results and effectiveness of logic
ID3 Algorithm
Results the project data, the results of training data, statistical data analysis results
ADO.NET
API
The training data performance database, data analysis and statistics database, management information database etc.

Figure 2 Overall Framework Model of Physical Training's Data Mining System

Decision Tree Construction Algorithm of Physical Training Based on Data Mining

Construction of Decision Tree Algorithm: Decision tree construction algorithm can be completed through training set T, therein T={<x, Cj>}, while x=(a1,a2,…，an) is a training example. It has n attributes, which are listed in the attribute list (A1, A2,... An), respectively. Here ai stands for the value of attribute Ai. Cj∈C=（C1,C2,…，Cn）is the classification result of X. The algorithm can be divided into the following steps:

Select attribute Ai from the attribute list to regard as the classification attribute. If the value of attribute Ai is Ki, T is divided into Ki of subsets, Ti,...,Tk, therein Tij={<x,C>|<x,c〉}. Moreover, attribute value A of X is the value of Ki. Delete attribute Ai out of the attribute list. For every Tij (1<j<Ki), make T=Tij, if the attribute list is non-null, return (1); otherwise, output. At present, there are comparatively mature decision tree methods, including ID3, C4.5, CART and SLIQ, etc.

The vaguer and more disordered in target classification training example set is, the higher its entropy is. The clearer and more ordered in target classification training example set is, the lower its entropy is. ID3 algorithm selects "the attribute that can conduct optimal classification on training example set from attribute list" by applying the principle that "the larger information earning (gain) has, the more beneficial the classification of training example will be". Information gain of an attribute results in the reduction of system gun, because the segment example of this attribute is utilized. Calculation of every attribute's information gain and comparison of information gain are key operations of ID3 algorithm.

The mathematical model can be described as follows: set E=Fi*F2*…*. Fn is n-dimension finite vector space, therein Fj is finite disperse symbol set. The element e=<V1,V2,...,Vn> in E is called as living example, therein Vj∈Fj,J=l,2,...,n. Set P and N to regard as two example sets of E and F, calling as positive example set and negative example set, respectively. Assume that the sizes of positive example set PE and negative example set NE in vector space E are P and N, respectively. Based on the following two hypotheses:

① in a correct decision tree of vector space E, classification probability of any example has consistent probability with E's positive and negative examples

② a decision tree can make correct judgment on a living example to obtain required information content (entropy of original set E):

$$E(E) = -\frac{P}{P+N}\log\frac{P}{P+N} - \frac{N}{P+N}\log\frac{N}{P+N} \qquad (1)$$

If regard Attribute A as the root of the decision tree, it possesses V values (Vi, V2...Vv). It divides E into V subsets (E1,E2.. .Ev). Assume that Pi positive example and Ni negative example in Ei , information entropy of subset E is E(Ei) :

$$E(E_i) = -\frac{P_i}{P_i + N_i}\log\frac{P_i}{P_i + N_i} - \frac{N_i}{P_i + N_i}\log\frac{N_i}{P_i + N_i} = \sum_{j=1}^{c}\frac{P_{ij}}{E_i}\log\frac{P_{ij}}{E_i} \qquad (2)$$

Consider the root of Attribute A as the information gun after classification (use expectation after classification of A) and regard it as E(A):

$$E(A) == \sum_{i=1}^{v}\frac{E_i}{E}E(E_i) \qquad (3)$$

Select Attribute A to minimize E (A) in formula2, and information gain will be increased with it.

Generation of Decision Tree Algorithm: Decision tree induction algorithm ID3 algorithm is described as follows:

//return a decision tree

Function ID3(R: a non-categorical attribute set, C: categorical attributes, S: a training set).

Begin

If S is null, return a single node with the value of Failure;

If S is consisted of records whose values have the same attribute value, return a single node with this value;

IF R is null, return a single node. Its value is the categorical attribute with the highest frequency in the S record; endow the value that has the maximal gain (D, S) in R to D;

Endow the value of Attribute D to {dj[j=l,2, 3,-, m};

Endow the value of subset S, which is consisted of Dj's records with the corresponding value of D, to {sj|1,2,3， m}; Return a tree; its root tab is D, and branch tabs are dl, d2, d3,…， dm;

Construct the following trees, respectively: ID3(R-{D},C, SI), ID3(R-{D}, C,S2),...,ID3(R-{D},C, Sm);End IDS;

Practice and Application of Physical Training Based on Data Mining

Attributes include student no., name, class, long jump, long-distance race, basketball, dash, hurdles, volleyball, shot, high jump and total points in line with students' examination performance database table. Adopt overall sampling methods to select student achievements of Class1, Class3 and Class 5 to be considered as training set, own a total of 161 records. Copy the performance records of three classes to the table of training example. Use SUM function to determine qualified number and unqualified number of scores of a subject in training set. The data is shown in Table 1:

Table 1 Training Set Data

	Long jump	Long-distance race	Basketball	Dash	Hurdles	Volleyball	Shot	High jump
Qualified Number	81	56	33	31 38	90	38	101	81
Unqualified Number	78	103	126	128	121	71	121	58

Use information gain to conduct attribute selection. It can draw a conclusion that long-distance race can be capable to distinguish the attribute of dash qualification in training example. Create a tree node and the subchain of this node, and every subchain represents the only value of selected attribute. Make use of EXCEL's selective function to display 4 records of unqualified long-distance race and qualified dash. It indicates that if long-distance race is unqualified, dash is unqualified basically. Its accuracy rate is (104-100)/104=96.2%. The performance of other three classes is regarded as testing set to examine the degree of accuracy for the generated decision tree, which can draw the following rules, as shown in Figure3. Conclusion 1: IF performance of students' long-distance race is unqualified, THEN, the performance of dash is generally unqualified. Conclusion 2: IF performance of students' long-distance race and hurdles is unqualified, THEN the performance of dash is unqualified. Conclusion 3: IF the performance of students' long-distance race and hurdles is qualified, THEN the performance of dash is qualified.

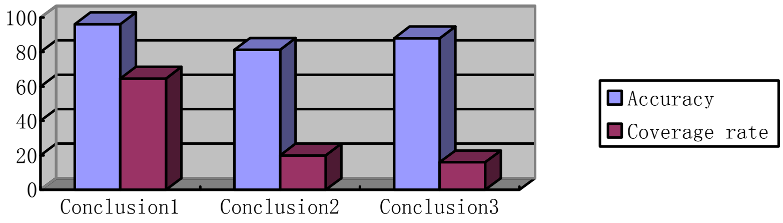

Figure 3 Structure Diagram

It can observe that learning degree of students' long-distance race will impact the learning effect of students' dash project directly. Learning of hurdles also has a certain influence on learning the dash. Therefore, when teachers conduct dash teaching, they should take students' long-distance race as the foundation. Students who have better degree of long-distance race and ordinary degree of dash should attach importance to learn hurdles.

Conclusions

The paper applies data mining to the system of physical training field, utilizes lots of experimental data to construct data warehouse and form multi-dimension data set of college students' physical quality, analyzes multi-dimension data set through data mining, and produces new knowledge rules to enrich knowledge base, so as to ensure accuracy of decisions. Moreover, the paper builds training model of data mining and ID3 decision algorithm. The experiment indicates that this model can better conduct statistic analysis on physical training data. With the continuous development of data mining and continuous in-depth study of sports scientific and technical personnel, no matter it is theoretical researches in data mining or practical research and development of data mining tools, all of them can bring great conveniences and considerable benefits to sports management decision and scientific research. Therefore, only to solve the above-mentioned problems, data mining can play a larger role on scientific development in sports field and own vaster potential for future development in sports field.

References

[1] Liao Shubing, Construction of College Physical Education Management Information-based System Based on Data Mining [J], 2011, 21(3):35-39.

[2] Chi Dianwei, Physical Training Decision Support System Based on Data Mining [J], Microcomputer Information, 2009, 25(4): 82-83.

[3] Xi Xianjie, Design and Implementation of Performance Management System [D], Zhejiang University of Technology, 2009, 13-16.

[4] Liang Xiexiong, Lei Nvhuan and Cao Changxiu, Research Progress on Modern Data Mining [J], Journal of Chongqing University, 2009, 27(3): 47-52.

[5] Tu Yun, Application of Data Mining in Sports Field[J], Journal of Wuhan Institute of Physical Education, 2012, 46(11): 27-30

[6] Huang Qian, Application Research of Data Mining in Physical Training Guidance[J], Journal of Guangzhou Sports University, 2009, 29(6):106-110.

[7] Song Xinshan et al., Application Research of Decision Tree Technology in Sports Teaching Quality Evaluation [J], Journal of Najing Sport Institute (natural science edition), 2009, 8(4):78-80.

[8] Yang Yaqin, Application Research of Data Mining in Personalized Service of Sports Distance Education [J], Journal of Beijing Sport University, 2006, 29(12):1614-1616.

[9] Li Xiaoling, Design and Implementation of School Sports and Health Management System [J], Journal of Ningxia University, 2009, 5(10):16-21.