

Community Life Sport Evaluation Based on C4.5 and Improved DRNN

Fei Gao¹ and RanLi²

¹Department of Physical Education, Anhui Jianzhu University, Hefei, China

²School of Management, Anhui Jianzhu University, Hefei China

¹tyff@ahjzu.edu.cn, ²ranran19780212@126.com

Abstract

With the development of our national sport consciousness, sports are a more and more important role in people's daily life. Community sports have been adopted by more and more people. But due to historical and economic reasons, community sport level in China is quite low. How to conduct an evaluation of the community life sport, not only is the community residents request, but also an important means to make corresponding improvement for a city builder. However, due to the mathematical algorithm has high complexity for the general case, applicants to environmental assessment is difficult. In this paper, introduce the C4.5 and improved DRNN as machine learning methods, and data mining the existing community life sport, to construct the evaluate model of community life sport, Not only can make up for the algorithm less shortcoming of current community sport research, but also can get the objective conclusion, to evaluate the mechanism of factors, and to give counsel for the development of community sport.

Keywords: community life sport, C4.5, the improved DRNN, Machine learning

1. Introduction

Since twenty-first century, China has constantly undertaking on Sport: qualify the football World Cup, the successful hold of the 2008 Beijing Olympic Games, breakthrough zero of winter Olympic gold medal. These memorable events not only enhance the morale of sport staff, but also inspired the enthusiasm of people in the whole country to participate in the sport[1]. We can say, in our history there is no time having so many sport achievements, and high enthusiasm of public participation in sport like today.

However, compared with the widespread sport participation enthusiasm, the construction of sport facilities in our community is relatively backward; the contradiction between supply and demand is very prominent. In our country, the old community generally has no special sport facilities, before the twentieth century 80's, even the school playground is paved with coal slag, not to mention the community sport facilities. Since the beginning of the new millennium, although China has in the largest civilize process in history, but because of the history and economy reason, the new community investment in sport facilities is still relatively small, and the corresponding, community life sport level in China is relatively low. At present, there is a variety of places for Chinese mass sport activities, these places are divided into sport guidance station, ordinary activities station, have fixed indoor venues and Sport Center (room, station) and sport team or group (Association) the five categories, The ordinary activities station's number is the most, accounted for 64.3%; Sport Center (room number, station) ranks second, accounting for 16.5%; Formal sport guidance station and sport activities of the association points are less, each accounted for 5%. Thus, mass sport activity standardization degree is low; the majority of activity is still in the non standardized state. We divided sport point scale into

large (more than 101), middle (30 ~100) and small (30 or less) three categories, among them medium-sized activities proportion is the largest, accounting for 48, 89%, Small and large activities were smaller, more than 101 large-scale activities accounted for 25.3%, small activities accounted for 26.29%, the reason may be related to site requirements and physical training atmosphere. Small site area`s requirements is not high, easy to manage, but it is difficult to form a good atmosphere of physical training; the large activities can form a good training atmosphere, but the more number of people required field is larger [2-6].

In 2013, because of the lack of community sport facilities the residents` contradictions can be seen everywhere on internet: Square dancers were suffering residents` water bombing, the square dance noise initiation murder case, competing to use community sport facilities and triggered disputes... we can say, the contradiction of national life sport demand and lack of community sport condition is cannot be ignored.

However, conflicting findings do not represent the solute of the conflict, community life sport are not the only system engineering, but also a sociological proposition, and it is not a short duration to enhance the level of community life sport. In 1978, UNESCO put forward "sport is one of the basic rights of all people". And the community sport is a typical life sport, is the basic way to realize "Life Sport" and "lifelong sport"[7].

Through the searching found, Chinese scholars have a larger proportion in the academic circles on this issue:

Somebody puts forward, to let sport become a part of city residents` living, it is necessary to study the relationship between the characteristics of city life and sport, make sport become citizens` living behavior [8].

Domestic scholar Lu Yuanzhen thought, the living sport is a kind of modern concepts of health and sport, is the rational sport behavior in the people`s daily life [9].

Tang Guojie thought: the community sport is a sociological category. It not only has the essential attribute of sport science, but also has the social attribute. To understand the evaluation problem of community sport should include social evaluation and physical evaluation two meanings. The social and community is subordination, the community and the community sport is a subordinate relationship. From the development of society and sport, social and sport is the subservient relationship. Sport is the development product of human society. On the other hand the development of sport promotes the continuous progress of human society [10].

Xiong Xuemei, analyze the cause of unbalance situation from the perspective of sociology, found that the distribution of social resources, population distribution, education and economic status is the root note of the unbalanced development of urban and rural sport [11].

Jing Yonggen, TianYupu *et al.* studied the balanced way to develop the urban and rural sport [12, 13].

Zhang Zhimei used the Delphi method, interview method, mathematical statistics, research methods such as literature, define the "Life Sport", "the lifestyle community of sport" concepts; and put forward the evaluation index system of sport [14-19].

From the above data is not difficult to see, although our discussion on community sport relatively richer than foreign, But most arguments are about macro statement and fact finding. The research results have the subjectivity of the author, lacking of mathematical support. However, due to the mathematical algorithms often have high complexity, and the requirement of normalization data is very high, introduction them to environmental assessment is very difficult. Therefore, based on the previous research, references C4.5-DRNN algorithm as a machine learning method, and mining the existing community life sport data to construct the evaluation model and then have an evaluation empirical community life sport in China. Not only can effectively ensure the objectivity of research results, but also can serve to find the main contradiction, to improve the construction of

community sport facilities.

The article is structured as follows:

The first section, the foreign and domestic research on community life sport are introduced,

In section second, the decision tree C4.5 algorithm theory and application

The third section, the relevant theory and application DRNN

The fourth section, constructs the evaluation model and empirical evaluation using the improved DRNN algorithm and the C4.5.

The research steps are showed in Figure 1.

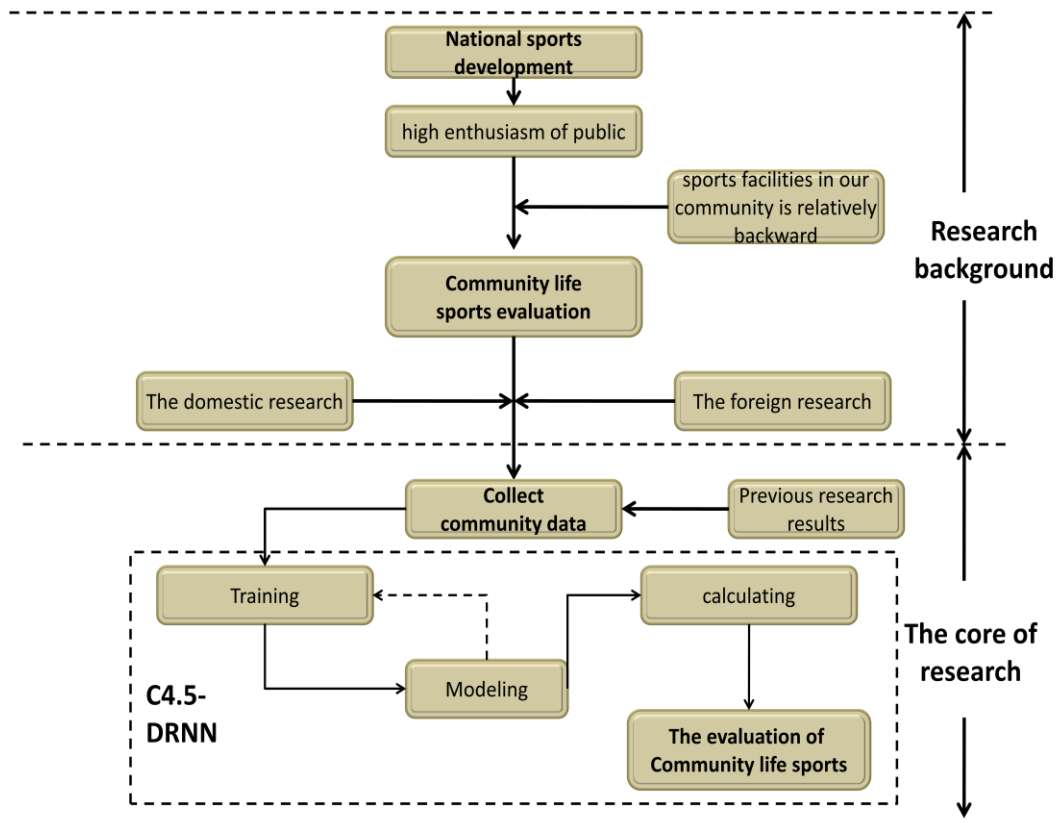


Figure 1. Research Step

2.C4.5 Algorithm

To achieve the purpose of forecasting, the decision tree classifies the data. The decision tree method according to the training set data to forming a decision tree; if the tree cannot give the correct classification of all objects, choose some exceptions added to the training set data, the process is repeated until the correct decision set formation. Decision tree represents the tree structure of the decision set.

The decision tree is made by the decision node, branches and leaves. The top node as the root node of the decision tree, each branch is a new decision node, each decision node representing a problem or decision, each leaf node represents a possible classification outcome. In the top to down traversal of the decision tree, each node will encounter a test, the different test output of each node problems leads to different branch, finally reach a leaf node, this process is the using decision tree classification process. The decision tree process is showed as Figure 2:

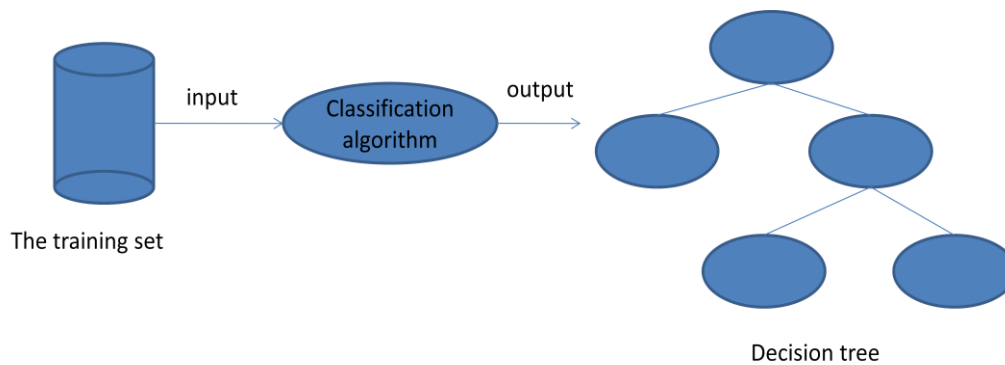


Figure 2. Decision Tree Process

Given a database $D = \{t_1, t_2, \dots, t_n\} \quad t \in D$, set of classes defined $C = \{C_1, C_2, \dots, C_m\}$ classification problems from the database to set the mapping from $f : D \rightarrow C$,

The database tuples t_i assigned to a category C_j ,

$$C_j = \{t_i | f(t_i) = C_j, 1 \leq i \leq n, t \in D\}$$

The ID3 algorithm is first proposed by Quinlan. The algorithm is built on the information theory, the information entropy and information gain degree as the standard, to achieve the classification of data. The followings are some basic concepts of information theory:

Definition 1: If there are n identical probability message, the probability P of each message is $\frac{1}{n}$, information content of a message transfer of $\log_2\left(\frac{1}{n}\right)$

Definition 2: If there are n messages, the given probability distribution $P = (p_1, p_2, \dots, p_n)$, the amount of information from the distributed transfer called P entropy, note

$$I(P) = -\sum_{i=1}^n p_i \log_2(p_i)$$

Definition 3:

If a collection of records of T according to the Category attribute value is divided into independent C_1, C_2, \dots, C_n , Where P is the probability distribution of C_1, C_2, \dots, C_n ,

$$P = \left(\frac{|C_1|}{|T|} \dots \frac{|C_k|}{|T|}\right)$$

Definition 4:

According to the non-categorical values X of T into the set T_1, T_2, \dots, T_n , determine element class information in T can be determined by a weighted average of T_i , namely $\inf o(T_i)$ the weighted average:

$$\inf o(X, T) = (i=1 \text{ to } n \text{ sum}) \left(\left(\frac{|T_i|}{|T|} \right) \inf o(T_i) \right)$$

Definition 5: information gain is the difference between the two of the amount of information, one of the amount of information is the amount of information required to determine an element of T , another information is the amount of information in the X property has been obtained when the value of the T element, information gain degree formula:

$$Gain(X, T) = \inf o(T) - \inf o(X, T)$$

Its core is the selection of attributes of each node in the decision tree, using the information gain as the attribute selection criteria, when testing, in each non leaf node can

obtain the biggest example information about the tested categories. The information gain of ID3 algorithm is adopted to select attributes are required to test. It is built on the concept of entropy in information theory.

ID3 algorithm has solved many problems in the practical application, for the non incremental learning tasks, ID3 algorithm is a good choice, but for incremental learning tasks, because ID3 is not incremental training samples, makes every increased case must abandon the original decision tree, restructure the new decision tree, this caused a great deal of overhead to the system. Then the ID3 algorithm is extended to C4.5 algorithm by Quinlan himself [20-22].

C4.5 algorithm not only has the function of ID3 algorithm, also increases the corresponding function, the increased the function include: information gain replaced with gain ratio; combined with continuous valued attributes; the missing attribute values in the training set; using different pruning techniques; k iteration of cross validation; produce the corresponding rules. The key is the C4.5 algorithm selected the highest information gain rate attributes as the test attribute

A training set T according to the discrete properties x of n different values, divided into T_1, T_2, \dots, T_n n subset of T_i using the x classification information gain rate:

$$Gain_ratio(x) = Gain(x) / Split_inf o(x)$$

The $Split_inf o(x)$:

$$Split_inf o(x) = - \sum_{i=1}^n ((|T_i|/|T|) \times \log_2(|T_i|/|T|))$$

Decision tree based on using the gain ratio of $Gain_ratio(x)$ is stronger than the decision tree building with $Gain(x)$. Gain ratio is the ratio of the maximum of the selection of the given attributes.

ID3 algorithm and C4.5 algorithm are very effective for relatively small data sets. If applied to large databases, its effectiveness and scalability have become the focus of attention of the algorithm. Despite the fact that C4.5 algorithm is improved for the data missing attribute values in the training set and the training set continuous attributes values, but also pay a memory and time cost. The decision tree algorithm in time complexity and space complexity is to be studied.

The ideal decision tree can be split into 3 types: the number of leaf nodes at least; the leaf nodes of the minimum depth; the number of leaf nodes and leaf nodes at least and the minimum depth. Decision tree is good or bad, not only affects the efficiency of the classification, but also influences the accuracy of classification. People want to have the ideal solution must seek various heuristic methods [23-27].

3. The Amproved DRNN Algorithm

3.1 Diagonal Recurrent Neural Networks

Diagonal Recurrent Neural Network is simplified FRNN connected to a recursive neural network. This paper is used for DRNN network model of multi input single output. Compared to the standard feedforward network, the difference is the hidden layer node DRNN with self feedback. Compared to the fully connected FRNN, there is no mutual exchange of information between hidden layer nodes. The model was greatly simplified.

The input vector and output vector dynamic mapping of DRNN expression:

$$\begin{aligned}
 (1) \quad O(k) &= \sum_j W_j^o X_j(k) \\
 (2) \quad X_j(k) &= f(S_j(k)) \\
 (3) \quad S_j(k) &= W_j^h X_j(k-1) + \sum_i W_{ij}^i I_i(k)
 \end{aligned}$$

Among them, $I_i(k)$ is DRNN's i input. $S_j(k)$ is the sum of the input j recursive neurons. $X_j(k)$ is the output of the j recursive neurons. The $O(k)$ is the output of the network output node. W^i, W^h, W^o are the input layer to the hidden layer weights of the hidden layer, self feedback, and hidden layer to output layer. Incentive function of $f(\cdot)$ is typical sigmoid function (hyperbolic) $f(x) = \frac{1-e^{-x}}{1+e^{-x}}$. The learning algorithm is using the steepest descent method, so the network has the problem of local minima. In order not to make the network into a local minimum, we use a genetic algorithm to optimize the weights of DRNN network, to find an optimal solution.

3.2 Use Genetic Algorithm to Optimizing the Weights of DRNN Neural Network

Genetic algorithm is rooted in biology in Darwin's theory of evolution, proposed in 1975 by Holland. The main characteristic of the genetic algorithm is simulated in the simulation by chromosome and chromosome evolution operation as a machine learning method, then according to the objective function to get a fitness function, to obtain the most adapted chromosomes as the optimal solution. The concrete can be described as Figure 3:

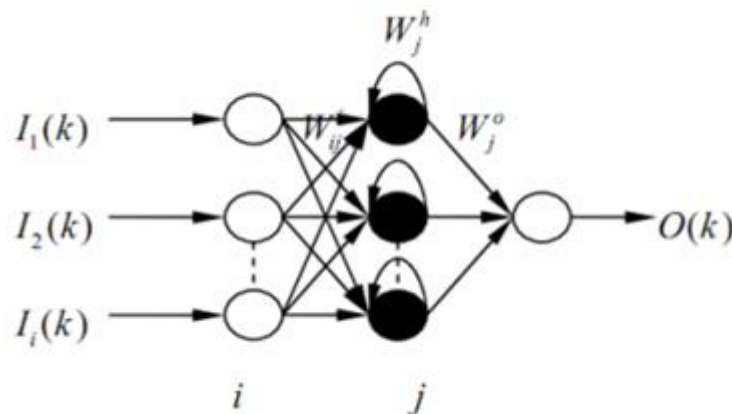


Figure 3. Diagonal Recurrent Neural Network Structure

Step1. Initialization. Sort the initial weights of each group according to the W^i, W^h, W^o order, i is an input layer of the i node, j is the j node of the hidden layer.

Step2. Randomly generates the initial population of solutions.

Step3. Calculate the fitness of each individual group. Fitness function $Fitness = \frac{1}{N} \sum_i (\tilde{O}(i) - O(i))$

Among them, N is the sample number of training samples; $\tilde{O}(i)$ is expectation of the i samples of the network output; $O(i)$ is the actual i samples of the network output value.

Step4.1 selection, according to fitness high principle, select individual from the old group to formulate group A ,group A has the same size of old, old group of individuals may appear multiple times in group A or not;

Step4.2 Cross, according to the crossover probability, in group A repeatedly and randomly selected individuals to obtain group A and group B ;

Step4.3 Mutation, the mutation probability, mutation of each individual group B , gets new groups.

Step5. If the group performance has met the requirements, or has reached the maximum number of iterations, the algorithm stops. The highest adapting degree of individual as a result of the output group; otherwise it returns Step3.

The complexity of the algorithm is very high, the manual calculation is difficult to achieve, and the actual computation process mainly relies on the computer [28-31].

4. Community Life Sport Evaluation is Based on C4.5and Improved DRNN

4.1. The Selection of Evaluation Index

In the research of the community sport,Professor Zhang Zhimei of Zhongzhou University organized the community sport experts, scholars fromvarious regions of the countryin 2012 April, make 4 rounds of the survey to identify community sport living level evaluation index system, get the wide attention of the academic circles. Through the deeper research, the system has been able to mature with the actual work.The evaluation index system is as Table 1:

Table 1. Evaluation Index System

Project	Content
Public service	The basic facilities, physical education, sport organization
Socialization	Spontaneous sport organizations, sport venues, backbone, funds and equipment to obtain the socialization degree
Crowd behavior	The number of sport population, the population quality of community sport, sport participation behavior and effect
The family sport	Family sport behavior characteristics, family sport consumption, sport family number

4.2 The Experimental Design

In the beginning, first through various channels to collect 20 communities' data as training samples, the 20 district not only includes national community sport demonstration area, but also contains other recognized community sport level in poor areas. Range:
 $D(\text{public service}) = \{\text{poor}, \text{medium}, \text{good}\}$, quantified as $\{0.2, 0.5, 0.8\}$
 ; $D(\text{social}) = \{\text{poor}, \text{medium}, \text{good}\}$, quantified as $\{0.2, 0.5, 0.8\}$
 ; $D(\text{behavior}) = \{\text{poor}, \text{medium}, \text{good}\}$, quantified as $\{0.2, 0.5, 0.8\}$
 ; $D(\text{family sports}) = \{\text{poor}, \text{medium}, \text{good}\}$, quantified as $\{0.2, 0.5, 0.8\}$
 , quantization; classification results set for $\{N, P\}$, quantitative $\{0.2, 0.8\}$
 , these communities' data as shown in the following Table 2:

Table 2. Communities' Data for Training

Public service	Socialization	Crowd behavior	The family sport	N/P
0.2	0.8	0.5	0.2	0.2
0.5	0.8	0.8	0.8	0.8
0.2	0.2	0.8	0.2	0.2
0.2	0.8	0.5	0.2	0.8
0.8	0.2	0.8	0.2	0.8
0.2	0.8	0.2	0.5	0.2
0.5	0.2	0.2	0.8	0.2
0.5	0.2	0.2	0.2	0.2
0.5	0.8	0.8	0.8	0.8
0.2	0.2	0.5	0.2	0.2
0.8	0.8	0.8	0.5	0.8
0.5	0.2	0.2	0.2	0.2
0.2	0.5	0.2	0.2	0.2
0.5	0.2	0.2	0.2	0.2
0.8	0.8	0.8	0.8	0.8
0.2	0.2	0.5	0.2	0.2
0.5	0.2	0.2	0.2	0.2
0.2	0.2	0.8	0.2	0.2
0.8	0.5	0.5	0.8	0.8
0.5	0.8	0.8	0.8	0.8
0.5	0.2	0.2	0.8	0.8

Use decision software, training through the improved DRNN and C4.5 algorithm can get the decision tree as Figure 4:

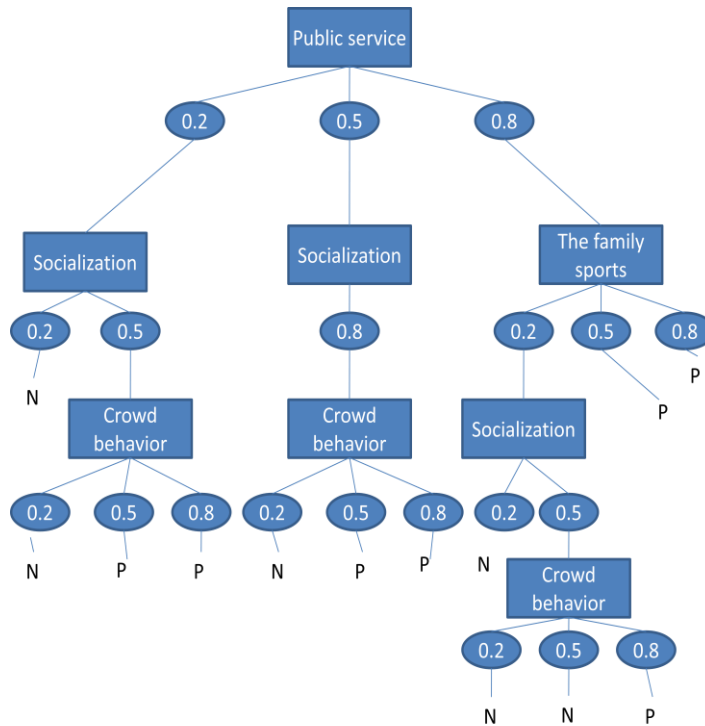


Figure 4. The Decision Tree

The model obtained, and then according to it calculates the community living sport level in 15 communities.

The experimental communities` data are shown in the following Table 3:

Table 3. The Experimental Communities Data

Public service	Socialization	Crowd behavior	The family sport
0.2	0.2	0.5	0.2
0.2	0.5	0.2	0.8
0.2	0.5	0.5	0.2
0.2	0.5	0.8	0.2
0.5	0.8	0.2	0.2
0.5	0.8	0.5	0.5
0.5	0.8	0.8	0.8
0.5	0.8	0.5	0.2
0.8	0.2	0.2	0.8
0.8	0.5	0.5	0.2
0.8	0.8	0.8	0.5

0.8	0.2	0.2	0.2
0.8	0.2	0.5	0.5
0.5	0.2	0.2	0.2
0.8	0.8	0.8	0.8

The calculation results as follow Table 4:

Table 4. Experimental Communities Living Sport Level

experimental communities	good or not
experimental community 1	0.2
experimental community 2	0.2
experimental community 3	0.8
experimental community 4	0.8
experimental community 5	0.2
experimental community 6	0.8
experimental community 7	0.8
experimental community 8	0.2
experimental community 9	0.2
experimental community 10	0.8
experimental community 11	0.8
experimental community 12	0.2
experimental community 13	0.2
experimental community 14	0.2
experimental community 15	0.8

7. Conclusion

In this paper, according to the improved DRNN algorithm and C4.5 algorithm, 15 areas of community living sport level were evaluated, and also understand the effect degree of every evaluation index in the evaluation process.

But this paper also has some shortcomings:

- (1) Selected 20 representative cell data as training data model. The quantity is small.
- (2) The 15 District data were collected through investigation, expert scoring method; collect means has certain artificiality;
- (3) The experimental results also have historical limitations.

These questions will be resolved in future studies.

Acknowledgment

This paper belongs to the project of “the research of multi-source information service system under the cloud computing environment ”supported by Chinese National Natural Science Foundation (No. 71131002).

References

- [1] X. M. Xiang, “Introduction to the physical environment”, Beijing Sport University press, (2003).
- [2] W. Kaizhen, “Research on the status qua of Beijing city community sport”, sport science, no. 5, (1994).
- [3] L. Depei, “Study our country city mass sport development trend to the present stage”, sport science, no. 3, (1994).
- [4] W. Kaizhen, “Current situation and development trend of China's city community sport”, sport science, no. 3, (1997).
- [5] L. Jiangnan, “Probing into the present situation of social sport in the Pearl River Delta and its social factors”, sport science, no. 3, (1995).
- [6] L. Shuting, “Guangzhou city community sport development mode”, sport science, no. 6, (1997).
- [7] W. Huan, “campaign on the home front”, Beijing: People's sport press, (2006), p. 47.
- [8] L. Jianguo, “Life sport and city sport life”, sport scientific research, vol. 27, no. 4, (2006), pp. 11-13.
- [9] L. Y. Zhen, “Sport belongs to life”, sport scientific research, vol. 27, no. 4, (2006), pp. 1-3.
- [10] T. Guojie, “Assessment of community sport sociology paradigm”, Journal of Hangzhou Normal University (Natural Science Edition), no. 5, (2006).
- [11] X. X. Mei, “Research on sport development imbalance between urban and rural in China”, Journal of Southwestern Normal University (Natural Science Edition), no. 8, (2010).
- [12] J. Y. Gen and X. X. Si, “The rural sport management and the farmer sport activities”, Beijing: China Society Press, (2006).
- [1] X. M. Xiang, “Introduction to the physical environment”, Beijing Sport University press, (2003).
- [2] W. Kaizhen, “Research on the status qua of Beijing city community sport”, sport science, no. 5, (1994).
- [3] L. Depei, “Study our country city mass sport development trend to the present stage”, sport science, no. 3, (1994).
- [4] W. Kaizhen, “Current situation and development trend of China's city community sport”, sport science, no. 3, (1997).
- [5] L. Jiangnan, “Probing into the present situation of social sport in the Pearl River Delta and its social factors”, sport science, no. 3, (1995).
- [6] L. Shuting, “Guangzhou city community sport development mode”, sport science, no. 6, (1997).
- [7] W. Huan, “campaign on the home front”, Beijing: People's sport press, (2006), p. 47.
- [8] L. Jianguo, “Life sport and city sport life”, sport scientific research, vol. 27, no. 4, (2006), pp. 11-13.
- [9] L. Y. Zhen, “Sport belongs to life”, sport scientific research, vol. 27, no. 4, (2006), pp. 1-3.
- [10] T. Guojie, “Assessment of community sport sociology paradigm”, Journal of Hangzhou Normal University (Natural Science Edition), no. 5, (2006).
- [11] X. X. Mei, “Research on sport development imbalance between urban and rural in China”, Journal of Southwestern Normal University (Natural Science Edition), no. 8, (2010).
- [12] J. Y. Gen and X. X. Si, “The rural sport management and the farmer sport activities”, Beijing: China Society Press, (2006).
- [13] T. Yupu, Y. X. Ming and L. K. Yun, “Unified development strategy of mass sport in urban and rural China”, Journal of physical education, vol. 15, no. 1, (2008), p. 10.
- [14] Z. Z. Mei, “The lifestyle community of sport evaluation index system research”, Journal of Beijing Sport University, no. 4, (2012).
- [15] Z. D. Gao, Y. D. Ming, “Study on the index system of capital Sport Modernization”, Journal of Beijing Sport University, vol. 5, no. 5, (2007), p. 581.
- [16] Beijing Municipal Bureau of statistics, Research on social science and technology development index system of Beijing City Social, Beijing Statistical Bureau, (2004).
- [17] L. X. Hua, “China modern sport and sport modernization”, Journal of physical education, vol. 9, no. 5, (2002), pp. 20-22.
- [18] S. Q. Zhu, “sport practical fuzzy mathematics”, Beijing: People's sport press, (1990).
- [19] C. Y. Hua, “Study on the evaluation index system of city construction”, Nanjing social science, no. 5, (2004), pp. 85-86.
- [21] X. W. Yan, “Application of ID3 algorithm in the analysis of English Achievements”, Journal of Liuzhou Vocational and Technical College, no. 11, (2011).
- [22] W. X. Wei, J. Y. Ming, “Analysis and improvement of ID3 algorithm of decision tree [J] computer engineering and design, no. 32, (2011).

- [23] N. W. Ying, "Application of decision tree ID3 algorithm in the university management information", exam week, no. 56, (2011).
- [24] J. R. Quinlan, "Induction of Decision Trees", Machine Learning, vol. 1, no. 1, (1986), pp. 81- 106.
- [25] J. R. Quinlan, "C4.5: Programs for Machine Learning", San Mateo, California: Morgan Kaufmann, (1993).
- [26] S. Ruggieri, "Efficient C4.5", IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 2, (2002), pp. 438- 444.
- [27] J. R. Quinlan, "Bagging, Booting and C4.5 in proc of 13th National Conference on Artificial Intelligence Portland", (1996), pp. 725 – 730.
- [28] E. Swere and D. J. Mulvaney, "Robot navigation using decision trees", Loughborougk, UK: Electronic Systems and Control Division Research, (2003).
- [29] T. B. Jun, "Soft instrument research of Polyester viscosity based on DRNN network", Beijing: Master's thesis of Beijing University of Chemical Technology, (2001).
- [30] Z. G. Sheng, Y. Ling and X. X. Ji, "DRNN based on improved ant colony algorithm and its application to dynamical system identification", China technology online, (2009).
- [31] G. Na, "A decision tree based on improved DRNN network construction method", Journal of Lanzhou University, (2011).

Authors



FeiGao. She received her master's degree in Physical education, with research direction being Physical education and training. Now she is a lecturer of Anhui Jianzhu University.



RanLi. He received his master's degree in Master of management, Ph. D. student. His research direction is environmental cost assessment and information management. Now he is an associate professor of Anhui Jianzhu University.