MATEC Web of Conferences **25**, 03009 (2015)
DOI: 10.1051/matecconf/20152503009

The Evaluation on Data Mining Methods of Horizontal Bar Training Based on BP Neural Network

Yanhui Zhang
Teaching and Research Office of Physical Education, Department of Public Basic Courses, Langfang Health Vocational College, Langfang, Hebei, China

Shaoqing Liu
Langfang Health Vocational College, Langfang, Hebei, China

ABSTRACT: With the rapid development of science and technology, data analysis has become an indispensable part of people's work and life. Horizontal bar training has multiple categories. It is an emphasis for the research of related workers that categories of the training and match should be reduced. The application of data mining methods is discussed based on the problem of reducing categories of horizontal bar training. The BP neural network is applied to the cluster analysis and the principal component analysis, which are used to evaluate horizontal bar training. Two kinds of data mining methods are analyzed from two aspects, namely the operational convenience of data mining and the rationality of results. It turns out that the principal component analysis is more suitable for data processing of horizontal bar training.

Keywords: data mining; BP neural network; cluster analysis; principal component analysis

1 INTRODUCTION

With the advent of the big data era, data mining methods are now playing a more and more important part in people's work and life, especially in sports, finance and other fields. Analyses on big data problems and the pursuit of relevant instructions have become effective ways and means to solve the problem.

In 2012, Yun Xu studied various applications of data mining in sports, drawing a conclusion that the application problem of data mining needs further studies [1]. For example, tactical analyses of real-time games or after games, the monitoring of national physique and health, and other data analyses. This research mainly includes three key points. First, the establishment of a sports data platform is the basis of data mining, which is beneficial to data communication and resource sharing among sports workers. Second, based on existing resources, cross-technological tools used to develop various data mining methods are essential means to realize the convenience of data mining with a view to the applicability and the universality of data mining. At last, the integration with technologies other than data mining is an important way of promoting the value of data mining. For example, the integration of the simulation system of various sports events and data mining technology can further realize simulation studies of sports.

In 2013, Xiangyang Xie illustrated relevant theoretical problems of data mining, such as the concept of data mining, applications of classification and prediction in data mining, traditional ways of data mining and the general process of data mining [2]. Meanwhile, the author carried out a series of discussions on data mining in terms of sports data analyses. In addition to relevant theories, the author also indicated important applications of data mining in various sports events with basic data of football, basketball and track and field trainings as analytical objects, especially in the aspect of data analysis. In physical education, data mining plays an irreplaceable role in creating humanized education classrooms. Finally, the author concluded that the data mining technology plays an important part in promoting sports events and physical education.

In 2014, Xinhui Zhao organized relevant documents on data mining's applications in sports research, which were categorized based on the knowledge of mathematical statistics. The six major categories include management, match, training, teaching, summary and others [3]. Results show that there are problems in the results of data analyses due to the fact that the main research focus lies in the theoretical analysis while studies on the database establishment and applications are quite less. For this reason, the author pointed out that it is still an important research direction that the data mining technology should be truly applied to tactical analyses and instructions of sports games.

With the difficulty of horizontal bar movements as the example, this paper compares two methods of data mining from the aspect of reducing categories of horizontal bar movements so as to provide instructions for practical trainings of horizontal bar.

Table 1. Groups of horizontal bar movements in the 29th Olympic Games

Athletes	I	II	III	IV	V
1	1D	2D1E	3D	1D1E	1E
2	2C	1D1E1F1G	1B	1B1E	1F
3	2C	1C2D1F	1E	1D1E	1E
4	2C1D	2D1F1G	1B	1D	1E
5	2C	1D1F	1B1C1D	2D	1E
6	1C	1D1F	2C2D	1B1E	1E
7	1C1D	1D1G	1D	1B1C1D1E	1E
8	1C1D	1F	1D1E	1B1C1D1E	1D

Table 2. Groups of horizontal bar movements in the 30th Olympic Games

Athletes	I	II	III	IV	V
1	1C	1D1E1F1G	1D1E	1D1E	1E
2	1C1D	2D1E1F	1E	2D	1E
3	1D	3D1E	1D1E	1D1E	1E
4	1D	3D1E	1D1E	1D1E	1E
5	1C	2D1E1F	1D	3D	1E
6	2B1D	1D1E1F1G	1B	1D	1E
7	1D	2D1E	1D1E	1C2D	1E
8	1C2D	3D1F	1D	1D	1E

Table 3. Groups of horizontal bar movements in the 29th Olympic Games

Athletes	I	II	III	IV	V
1	14	2415	34	1415	15
2	23	14151617	12	1215	16
3	23	132416	15	1415	15
4	2314	241617	12	14	15
5	23	1416	121314	24	15
6	13	1416	2324	1215	15
7	1314	1417	14	12131415	15
8	1314	16	1415	12131415	14

Table 4. Groups of horizontal bar movements in the 30th Olympic Games

Athletes	I	II	III	IV	V
1	13	14151617	1415	1415	15
2	1314	241516	15	24	15
3	14	3415	1415	1415	15
4	14	3415	1415	1415	15
5	13	241516	14	34	15
6	2214	14151617	12	14	15
7	14	2415	1415	1324	15
8	1324	3416	14	14	15

2 GRADING OF HORIZONTAL BAR MOVEMENTS

Horizontal bar is an important sports match of the Olympic Games. The match is divided into five groups. Movement data of the horizontal bar finals in the 29th and 30th Olympic Games is taken as the example and data mining methods are studied, as shown in Table 1 and Table 2.

Data in Tables 1 and 2 are processed for the convenience of the cluster analysis and the principal component analysis. A–G are replaced with 1–6. Results are shown in Tables 3 and 4.

3 CLUSTER ANALYSIS

The cluster analysis is also called clustering analysis [5]. It is a multivariate statistical method for classification problems, the essence of which is the collection of similar elements. The way of combining qualitative description analysis and quantitative data is usually required if elements with strong correlations are divided into a category. Generally, multiple indicators are combined as a representative indicator. The representative indicator can be selected by judging differences of indicators through data analysis. Classification problems exist everywhere in our daily life and study. So cluster analysis has become a necessary way of data processing. It covers quick cluster, hierarchical

clustering and two-step cluster. The hierarchical clustering is adopted in this paper.

3.1 *Idea of cluster analysis*

Suppose there are *n* samples X_1, X_2, … , X_n in set *G*. First, each sample is a single individual. Calculate the distance between two individuals and the nearest two are clustered together. Second, calculate the distance between two new clusters and the nearest two variables are clustered as one until all variables are clustered as one variable.

3.1.1 *Preparations of cluster analysis*

1. Determine clustering types. The sample cluster is type *Q* and the variable cluster is type *R*.

2. Data preprocessing, such as data conversion.

3. The purpose of cluster analysis is the classification samples. It is very necessary to explore mutual relations of all kinds of samples. There are generally two methods:

(1) Calculate the similarity between two samples. If a variety of properties are similar, the similarity is higher which will not exceed 1.

(2) Calculate the distance between two points and a coordinate point represents a sample. Calculate the coordinate distance of each point and points in near distance are regarded as one type.

4. Calculate the distance matrix or the similarity coefficient matrix *D*.

3.1.2 *General steps of cluster analysis (Q type classification)*

(1) Each sample is an individual class, $G_i = \{X_i\}\ \ (i = 1,2,3\ldots,\ n)$.

(2) Construct the matrix *D* according to the calculated distance or the similarity. Meanwhile, find the smallest D_{ij} and combines G_i, G_j as one class $G_r = \{G_i, G_j\}$.

(3) Recalculate the distance between two classes to get the new matrix *D*.

(4) Repeat step (2) until all matrixes is clustered as one class.

There are eight methods of cluster analysis, namely sin, com, med, cen, ave, fle and ward.

3.2 *Results of cluster analysis*

Carry out a cluster analysis on data in Table 3. Results are shown in Figure 1. And carry out another cluster analysis on data in Table 4 in order to ensure the accuracy of the data mining method. Results are similar to those in Figure 1.

It can be seen from Figure 1 that group 3 and group 5 are the most similar, followed by group 2 and group 3. Group 2 and group 4 are less similar to each other. Group 1 is listed individually.

4 PRINCIPAL COMPONENT ANALYSIS

The main idea of principal component analysis is the dimensionality reduction of variable [6]. It is a statistical analysis method that converts multiple variables into a few main variables. It is commonly applied in data compression, system evaluation, regression analysis, weighted analysis and so on.

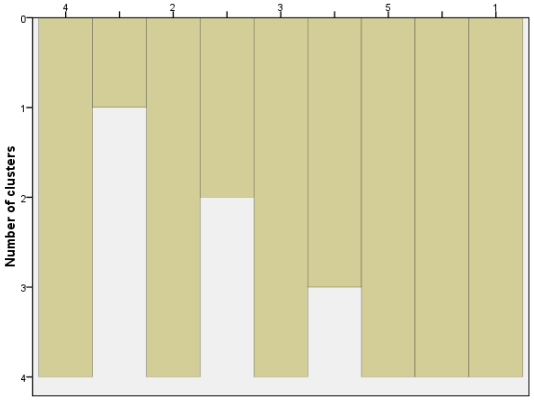

Figure 1. Cluster analysis results

4.1 *Concept of principal component analysis*

The main method of principal component analysis is to weaken multiple variables, only left variable that cannot be weakened. It actually means that many original variables are recombined into a set of independent variables so as to achieve the goal of reducing variables.

Suppose that there are *m* original indexes for principal component analysis which are $x_1, x_2, \cdots, x_m$. Now there are *n* samples and the observed value is $x_{ik}(i = 1,2,\cdots,n)$. After the standard deformation of $k = 1,2,\cdots,m$, x_k is changed into x_k^*. It is shown as follows:

$$x_k^* = \frac{x_k - \overline{x_k}}{s_k},\ k = 1,2,\cdots,m \tag{1}$$

Where, $\overline{x_k}$ and s_k are respectively the mean value and the standard deviation of x_k. The mean value of x_k^* is 0 and the standard deviation is 1.

The coefficient b_{kj} can be obtained according to the observed value x_{ik} of each original variable or the standardized x_{ik}^*. Establish a synthetic equation $z_j = \sum_k b_{kj} x^*_k$ about the standardized variable x_k^*. It is no doubt that the synthetic index equation z_j about x_k can be also established.

$$z_j = \sum_k \tilde{b}_{kj} x_k^* + a_j \quad (2)$$

There are two requirements about the determination of b_{kj}:

(1) Comprehensive indexes are mutually independent or uncorrelated.

(2) The amount of information of multiple samples reflected by comprehensive indexes equals to the eigenvalue of the corresponding eigenvector (the coefficient of comprehensive index). The contribution sum of the eigenvalue of selected comprehensive indexes is usually required to be larger than 80%.

4.2 *General steps of principal component analysis*

(1)Calculate x_k and $s_k(k, j = 1,2,\cdots,m)$ according to the observation data.

(2) The eigenvalue λ_j $(j = 1,2,\cdots,m)$ and multiple index values of each principal component can be obtained according to the correlation coefficient matrix R. The accumulative contribution rate is the criterion for determining the number of principal components *p*.

(3) m basic equations are presented below:

$$\begin{cases} r_{11}x_1^{(j)} + r_{12}x_2^{(j)} + \cdots + r_{1m}x_m^{(j)} = \lambda_j x_1^{(j)} \\ r_{21}x_1^{(j)} + r_{22}x_2^{(j)} + \cdots + r_{2m}x_m^{(j)} = \lambda_j x_2^{(j)} \\ \cdots \\ r_{m1}x_1^{(j)} + r_{m2}x_2^{(j)} + \cdots + r_{mm}x_m^{(j)} = \lambda_j x_m^{(j)} \end{cases} \quad (3)$$

Where, $j = 1,2,\cdots,m$.

Conduct the Schimidt orthogonalization and solve the basic equation of each λ_i, $x_1^{(j)}, x_2^{(j)}, \ldots, x_m^{(j)}$ $(j = 1,2,\cdots,m)$. Then order:

$$b_{kj} = \frac{x_k^{(j)}}{\sqrt{\sum_k \left(x_k^{(j)}\right)^2}} \quad (4)$$

The principal component $z_j = \sum_k b_{kj} x_k^*$ represented by $x_1^*, x_2^*, \ldots, x_m^*$ can be obtained. Or the principal component $z_j = \sum_k \tilde{b}_{kj} x_k^* + a_j$ represented by $x_1, x_2, \ldots, x_m$ can be obtained by substituting $x_k^* = \frac{x_k - \overline{x_k}}{s_k}$.

(4) Substitute observed values of x_1, x_2, …, x_m into the expression of the principal component to calculate the value of each component.

(5) Calculate correlation coefficients of the original indexes and the principal component, namely the factor loading, to indicate the principal component.

4.3 *Results of principal component analysis*

Carry out a cluster analysis according to data in Table 3 and the results are shown in Figure 2. Carry out another cluster analysis on data in Table 4 in order to ensure the accuracy of the data mining method. Results are similar to those in Figure 2:

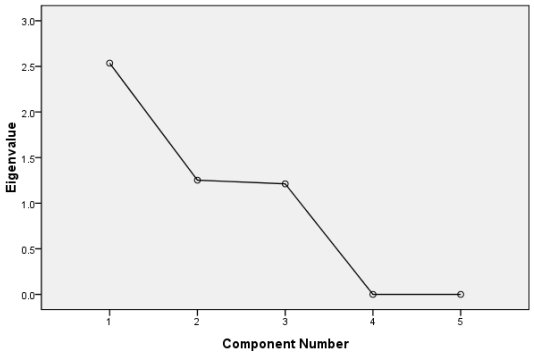

Figure 2. Scree plot

It can be seen from Figure 2 that groups with eigenvalue which is larger than 1 are groups 1, 2 and 3. Therefore, these three groups are principal components. According to the obtained results, groups 1, 2 and 3 can be particularly trained in practical trainings of horizontal bar.

The convenience of the data mining method and the rationality of the results are evaluated with BP neural network.

5 BP NEURAL NETWORK MODEL [7]

5.1 *Concept of neural network model*

Neural network model originates from neurobiology, the calculation process of which is similar to the reaction process of biological neurons. Neurotransmitters of different sources released by synapses have certain influences on membrane potential changes of same neurons. It can be seen that the ability of neurons of integrating information is spatially to integrate the input information of different sources on dendrite. Based on this ability, people created artificial neuron model by simulating the reaction process of neurons. Symbols in Figure 3 are described in Table 5.

The output form of the threshold θ_i is decided by $f[u_1]$ under the common effect of inputting $x_1, x_2, \cdots, x_n$. Graphs of two excitation functions are

presented in Figure 4. The second excitation function is adopted by the model of this paper.

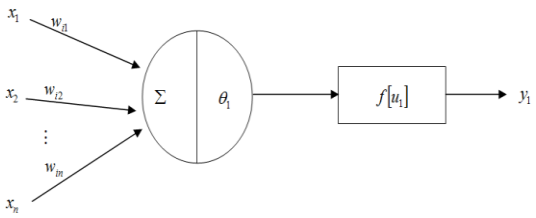

Figure 3. Schematic of mathematical models of neurons

Table 5. Descriptions of symbols in mathematical models

Symbols	Descriptions
$x_1, x_2, \cdots, x_n$	Input part of neurons, namely the information from the former grade
θ_i	Threshold of neurons
y_i	Output of neurons
$f[u_1]$	Excitation function

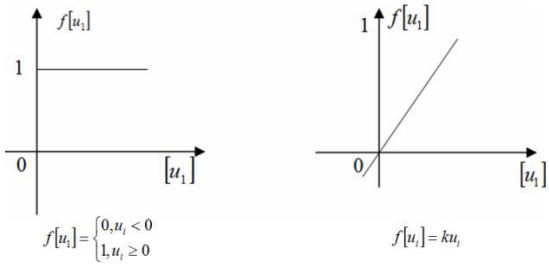

Figure 4. Typical excitation functions

Where,

$$u_i = \sum_j w_{ij} x_i - \theta_i \tag{5}$$

So,

$$y_i = f[u_i] = f\left(\sum_j w_{ij} - \theta_i\right) \tag{6}$$

Formula (2) is the integral mathematical model expression of a single neuron.

5.2 *Calculation steps of BP neural network model*

BP neural network is a kind of multilayer feed-forward network, the calculation method of which is the minimum mean square error. *Sigmoid* is used as the excitation function when the back propagation algorithm is applied in the multilayer feed-forward network. The network weight coefficient recursion w_{ij} is solved in the following steps. If there are n neurons in each layer, the i^{th} neuron of the k^{th} layer has n weight coefficients $w_{i1}, w_{i2}, \cdots, w_{in}$. In addition, select one more w_{jn+1} to represent θ_i. When the sample x is input, $x = (x_1, x_2, \cdots, x_n, 1)$.

1. Assignment of w_{ij}. Assign a relatively small nonzero random number to w_{ij} in each layer and $w_{jn+1} = -\theta_i$ in the meantime. This model is operated with Matlab, so the assignment is a random process of the computer. It is also because the same procedure code might have different results in different operational processes.

2. Input sample values $x = (x_1, x_2, \cdots, x_n, 1)$ and the corresponding expected output $y = (y_1, y_2, \cdots, y_n, 1)$.

3. Calculate the output of each layer. As for the output x_{ik} of the i^{th} neuron of the k^{th} layer,

$$y_i^k = f\left[u_i^k\right] \tag{7}$$

Where,

$$u_i^k = \sum_j w_{ij} x_j^{k-1} - \theta_i^k \tag{8}$$

In this formula, $x_{n+1}^{k-1} = 1$, $w_{i(n+1)} = -\theta$

4. Solve the calculation error $d_i^{\wedge}$ of each layer. As for the output layer, $k = m$ and

$$d_i^m = x_i^m \left(1 - x_i^m\right)\left(x_i^m - y_i^m\right) \tag{9}$$

As for other layers,

$$d_i^k = x_i^k \left(1 - x_i^k\right)\left(\sum_j w_{ij} x_j^{k-1} - \theta_i^k\right) \tag{10}$$

5. Revise w_{ii} and θ_i,

$$w_{ij}(t+1) = w_{ij}(t) - \eta d_i^k x_j^{k-1} \tag{11}$$

6. After the weight coefficient of each layer is solved, it is able to determine whether the requirements are met according to the established criteria. If the requirements are not met, go back to step3. Otherwise, end the calculation.

5.3 *Construction of the evaluation model*

Set the eigenvalue according to practical situations. A questionnaire survey is conducted for related sports workers who are required to set the evaluation standard for the convenience of data mining and the rationality of results in horizontal bar training. After the

processing, the investigation results are shown in Tables 6 and 7.

The analysis is carried out with the cluster analysis and principal component analysis on horizontal bar training data as the research object, results of which are shown in Figure 5.

Table 6. Parameters of excellent data mining methods

No.	1	2	3	4	5
Operating convenience	0.39	0.42	0.43	0.47	0.52
Rationality of results	0.76	0.84	0.77	0.66	0.87

Table 7. Parameters of inferior data mining methods

No.	1	2	3	4	5
Operating convenience	0.39	0.42	0.43	0.47	0.52
Rationality of results	0.76	0.84	0.77	0.66	0.87

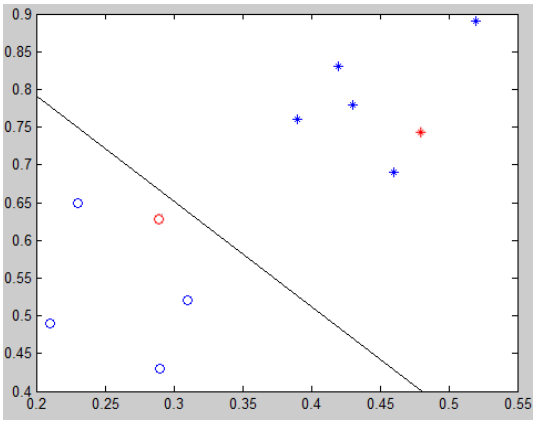

Figure 5. BP analysis results

Figure 5 is obtained by the programming of Matlab according to the calculation steps of BP neural network. In this figure, "*" stands for excellent data mining methods while "O" stands for inferior data mining methods. Red symbols represent data mining methods of horizontal bar training. It can be inferred from the figure that the principal component analysis is more suitable for data processing of horizontal bar training.

6 CONCLUSION

In this paper, BP neural network model is applied in the evaluation problems of horizontal bar training data. Based on the elaboration of calculation steps of cluster analysis and principal component analysis, this paper carries out data acquisition in the way of an actual survey and objectively reflects the evaluation situation from the aspect of data. Although, BP neural network is widely applied in evaluation problems in real life, neural network needs to estimate training errors reasonably. Once the error is not estimated reasonably, there might be incorrect calculation results.

REFERENCES

[1] Xu, Y. 2012. Applications of data mining in sports, *Journal of Wuhan Institute of Physical Education*, 46(11): 27-29.

[2] Xie, X.Y. 2013. Research and application of data mining in sports data analyses, *Contemporary Sports Technology*, 3(23): 9-10.

[3] Zhao, X.H., Shen, Y.W. & Zhang, J.F., et al. 2014. A review of the application status of data mining technology in sports science research, Zhongzhou Sports: Shaolin and Taiji, (7): 44-48.

[4] Tan, L. 2013. A Study on the Developing Trend of Men's Floor Exercise and Horizontal Bar in the World, Master's Thesis of Northwest Normal University.

[5] Zhou, Y.Z. 2010. *Mathematical Modeling*, Shanghai: Tongji University Press.

[6] Chen, S.K. 2010. SPSS Statistic Analysis: From Introduction to Proficiency. Beijing: Tsinghua University Press.

[7] Wang, X.Y., et al. 2010. *Mathematical Modeling and Mathematical Experiment*, Beijing: Science Press.