Computer Fraud and Security
ISSN (online): 1873-7056

Multiple Data Analysis of College Students' Physical Health Based on Big Data

Huizhong Zhang[1], Fanrong Meng[2,*] ,Qinyong Wang[1]
[1]Zhejiang College of Security Technology, Wenzhou, Zhejiang Province, China
[2]Wenzhou University,Wenzhou, Zhejiang Province, China
**Corresponding author*

Abstract:

College Students' physical health data has the characteristics of high dimension, and the amount of data is particularly large. Big data method and self-organizing feature mapping network (SOM) method have unique advantages and visualization characteristics for processing high-dimensional mass data, so they become important tools for pattern recognition and visualization analysis of big data. Taking Zhejiang College of Security Technology students' physical health data as an example, this paper analyzes the regional characteristics of influencing factors and explanatory factors of students' physical health with big data and visual SOM method. Results the analysis showed that body weight and BMI index had regional consistency, which were the most important factors affecting students' physical health, and were also the main explanatory variables of students' physical health status. The regional difference of physical health of girls is relatively large, while that of boys is small. The clustering characteristics of students' physical health indicators also have regional consistency. This paper demonstrates the rationality of big data method and self-organizing feature mapping network method in pattern recognition and visual analysis of physical health data. The results provide a certain reference value for the analysis of physical big data.

Keywords: Big data; SOM; data mining; physical health.

1. INTRODUCTION

Moderate physical exercise and abundant physical function are important guarantee of physical and mental health. Regular sports activities can help to avoid the problems of obesity, diabetes, hypertension and other stubborn diseases in the growth of teenagers, and to some extent, it is also conducive to improving and improving learning efficiency.Therefore, each country has set up corresponding university institutions, through the research on health promotion, comprehensively improve the physical and mental health of youth groups[1-2].

With the continuous advancement of urbanization, the economic level, urbanization complexity, social and cultural differences among different regions, as well as a series of social environment and ecological environment differences caused by them, make the production and living environment closely related to the healthy growth of teenagers have undergone fundamental changes, showing certain medical geographical characteristics, and the amount of physical health data [3-4].

The research method of data visualization has gradually become an important content of physical health research.For a long time, the research on students' physical health mainly focuses on the descriptive statistics and analysis of body shape, physical function and physical health and other index parameters, and fails to fully consider the influence of regional characteristics. There is a lack of in-depth and systematic research on the change mode and law of physical health reflected by big data of physical health.There are few qualitative research methods based on self-organizing map (SOM) to visually identify the pattern characteristics of physical health big data, and the empirical research on the influence factors and characteristics of physical health big data by principal component analysis (PCA) dimension reduction method are rare.Therefore, this study takes the data set of students' physical health in a university in Zhejiang Province as an example to explore the qualitative identification of the regional characteristics of students' physical health by SOM method from the perspective of regional characteristics of students' physical fitness, and discusses the regional characteristics of the main influencing factors and explanatory factors of students' physical health identified by visual PCA method.At the same time, in the context of big data, SOM and PCA methods will help to promote the research on pattern recognition and visualization of physical health data, which has certain scientific significance and practical value for the study of regional characteristics of students' physical health [5-7].

2. SOM PATTERN RECOGNITION AND VISUALIZATION OF STUDENTS' PHYSICAL CHARACTERISTICS

SOM is an artificial neural network with self-learning function. The neural network will be divided into different regions when it receives multi variable input from the outside, and different regions have different response characteristics to different variable patterns, and finally form a topological visual ordered graph.SOM can map the input signals of any dimension into a two-dimensional scatter diagram in topological sense. This analysis method is usually used to classify or identify the relationships and patterns among input variables.Because SOM can train and judge the input patterns by self-organization and realize the aggregation of neurons with the same function in spatial distribution, SOM is often used in the clustering analysis and qualitative research of big data in the field of Informatics because of its intuitive, visual and visual performance characteristics.In the research of SOM data pattern recognition and visualization, there are usually two methods to classify and analyze the data [8]:

(1) cluster analysis first, then carry out visual image calibration and projection, after projection, the same category of data for new image visualization expression;

(2) according to the mapping structure of the data itself for clustering and image visualization expression.For physical health data, the second method is often used because of the discreteness of its variables.The results of pattern recognition and visualization of physical characteristics of male and female students in different regions of sample areas by SOM show that there are differences in regional characteristics of physical measurement parameters between male and female students. The height and vital capacity of male students in Jinhua and Shaoxing areas are relatively the largest, and the boys in Wenzhou area are 50 m running and 1 000 m running. The results of m-running are relatively poor, and the change rules of boys and body mass index are similar, which shows that the change of boys is mainly affected by weight relative to height; the variation rules of boys' 50m running and standing long jump are consistent, which may be related to the short-term high explosive force required by the project; the height and vital capacity index values of girls in Jiaxing, Wenzhou and Yiwu are relatively large, with a value of 50The results of m-run and 800 m-run are relatively poor, and the change law of female students is similar to that of weight, and the change of female index is mainly affected by weight; the consistency of 50 m running and standing long jump is also reflected in girls. Logical architecture of big data analysis and processing platform is shown in Figure 1.

SOM has obvious advantages in dealing with multi-dimensional big data. Its method has unique visualization ability, which can directly reflect the change pattern of each parameter. By generating self-organization chart of each parameter, it can intuitively and qualitatively express the distribution characteristics of each parameter in a specific interval. In SOM analysis of this study, the Euclidean distance calculation, classification and visual expression of physical data characteristics are carried out based on U-matrix and K-means methods. Only 11 dimensions of clustering and qualitative analysis are carried out in different regions of the province, which reflects the general law of regional differences in students' physical health.In essence, in SOM analysis and calculation, the weight expression needs to use all the sample data, which contains a large amount of information. SOM can also map the input space sample pattern orderly to the output layer, and can map the high-dimensional data to the low-dimensional space to express clearly, and it is easy to find the rules.It should be noted that the number of selected variables, the number of samples and the number of clusters will affect the recognition effect of SOM, but for a given sample size data set, the mapping structure based on the information of the data itself can effectively distinguish the overall changes between different categories.

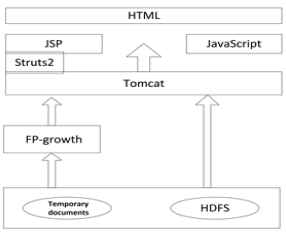

Fig. 1. Logical Architecture of Big Data Analysis and Processing Platform

Internationally, SOM is still in its infancy in pattern recognition and visual analysis of physical health data. For example, K. lagus and others have carried out pattern recognition and visualization of six physical fitness indexes and three symptom indicators of 371 researchers based on SOM, and discussed the relationship between physical fitness and symptoms; pellicer-chenoll and others used SOM to analyze the physical fitness, physical fitness, physical fitness, physical fitness and other indicators of 371 researchers based on SOM. The pattern recognition of body composition and academic performance is carried out, and the change characteristics of pattern in different time periods are discussed. It is proved that better physical fitness and better grades have similar pattern characteristics.

Association rules found a relationship between things and other transactions or interdependence. Assuming that $I=\{i_1,i_2,\ldots i_m\}$ is a collection and the related data task D is a collection of database transactions, in which each transaction T is a collection, and making $T \subseteq I$. Every transaction has an identifier TD. Assuming that A is a set of items, $A \subseteq T$. Association rules are the containing type of $A \Rightarrow B$, among them, $A \subset I$, $B \subset I$ and $A \bigcap B = \Phi$. The rules $A \Rightarrow B$ is in the transaction which sets up with support s, s is the percentage for the transaction contains $A \bigcup B$ in the D.

The LIPI algorithm through scanning data sets of frequency, then finding relevant data and finishing dig, its principle as follows:

Assuming that $I=\{i_1,i_2,\ldots i_n\}$ is a collection, which composed of different characteristics, the characteristics of each item as a constituted set of items. And the item set is not an empty set, but is a subset of the set of I, which can be expressed as$(x_1x_2\ldots x_m)$, every x_k is a term.

The sample variance and sample proportion of variance have established the following relationships [9]:

$$S^{*2} = \frac{S^2}{\mu^2} \quad (1)$$

Proof: by the definition, we have:

$$\begin{aligned} S^{*2} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\alpha_i - \overline{\mu}}{\overline{\mu}^2}\right)^2 \\ &= \frac{1}{(n-1)\overline{\mu}^2}\sum_{i=1}^{n}(\alpha_i - \overline{\mu})^2 \\ &= \frac{1}{\overline{\mu}^2}S^2 \end{aligned} \quad (2)$$

Consider delay, the L can be expressed as:

$$\mathcal{L}^0 = \begin{Vmatrix} C^0_{ijkl} & e^0_{kij} \\ e^{0T}_{ikl} & -\eta^0_{ik} \end{Vmatrix} \quad (3)$$

These functions can be expressed in the following form:

$$C(\mathrm{x}) = C^0 + C^1(\mathrm{x}),\ e(\mathrm{x}) = e^0 + e^1(\mathrm{x}),\ \eta(\mathrm{x}) = \eta^0 + \eta^1(\mathrm{x}),\ \rho(\mathrm{x}) = \rho_0 + \rho_1(\mathrm{x}) \quad (4)$$

The value with superscript of 1 represents the difference below:

$$C^1 = C - C^0, e^1 = e - e^0,$$

$$\eta^1 = \eta - \eta^0, \rho_1 = \rho - \rho_0 \quad (5)$$

The whole function can be simplified into the following integral equation set:

$$f(x,\omega) = f^{0}(x,\omega) + \int_{V} S(x-x')(\mathrm{L}^{1}F(\mathrm{y}') + \rho_1\omega^2 \mathbf{g}(R)\mathrm{T}_1 f(\mathrm{y}')]S(\mathrm{y}')\mathrm{d}\mathrm{y}' \quad (6)$$

In addition, we can introduce the abbreviated formula:

$$g(x,\omega) = \begin{Vmatrix} G_{ik}(x,\omega) & \gamma_i(x,\omega) \\ \gamma_k(x,\omega) & g(x,\omega) \end{Vmatrix}, \quad s(x,\omega) = \begin{Vmatrix} G_{ik,l}(x,\omega) & \gamma_{i,k}(x,\omega) \\ \gamma_{k,l}(x,\omega) & g_{,k}(x,\omega) \end{Vmatrix},$$

$$L^{1}(x,\omega) = \begin{Vmatrix} C^{1}_{ijkl} & e^{1}_{kij} \\ e^{1T}_{kij} & -\eta^{1}_{ik} \end{Vmatrix},$$

$$F(x,\omega) = \begin{Vmatrix} u_{(i,j)}(x,\omega) \\ \varphi_{,i}(x,\omega) \end{Vmatrix} \quad (7)$$

In these expression, $G_{ik}(x,\omega)$, $\gamma_i(x,\omega)$, $g(x,\omega)$ can be represented as:

$$g(x,\omega) = \frac{1}{(2\pi)^3}\int g(k,\omega)\exp(-ik\square x)dk \quad (8) \quad g(k,\omega) = \begin{Vmatrix} G_{ik}(k,\omega) & \gamma_i(k,\omega) \\ \gamma_k^{T}(k,\omega) & g(k,\omega) \end{Vmatrix} \quad (9)$$

Where $G_{ik} = (\Lambda_{ik} + \frac{1}{\lambda}h_i h_k^{T})^{-1}$,

$$g = -(\lambda + h_i^{T}\Lambda_{ij}^{-1}h_j)^{-1}, \quad \gamma_i = \frac{1}{\lambda}h_k^{T}G_{ki}. \quad (10)$$

3. Research Object and Method

According to the relevant requirements of the national student physical health standard, the parameters of student physical fitness test include six basic items: height, weight, body mass index, lung, 50m run and jump.Because of the coexistence of positive and negative values of sitting body flexion (bend) index, it is difficult to meet the relevant conditions of data standardization and ranking in PCA. Therefore, this index is not considered in PCA analysis.The specific indicators for boys included 1 000 m and pull-up, while those for girls included 800 m and 1 min sit up. Each physical fitness test index is determined in strict accordance with the standard method of "national student physical health standard". Logical architecture of big data analysis and processing platform is shown in Figure 2.

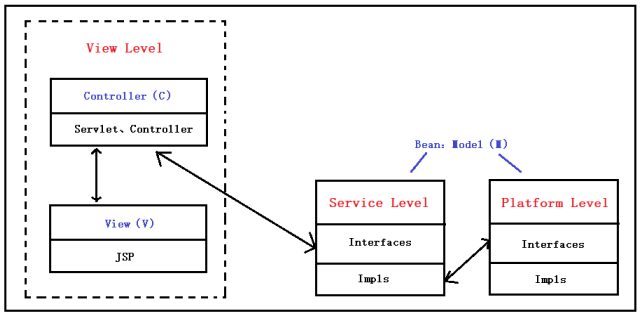

Fig. 2. Logical architecture of big data analysis and processing platform.

3.1 BASIC METHOD

SOM pattern recognition and visualization method SOM is an artificial neural network with self-learning function. It adopts competitive unsupervised network structure. The typical SOM network includes two layers: input layer and output layer. The meta structure network of SOM neural network is determined by the number of

input samples. In this study, the neuron type is hexagonal neuron, each neuron has six adjacent neurons, and the neural network is a rectangular network structure of 11 × 9 (the samples are divided into 11 regions and 9 physical fitness indicators). After determining the structure and size of the neural network, the neurons of each node are assigned an initial weight. There are two ways of SOM initialization: random initialization and linear initialization. Random initialization is to assign a small weight to each neuron vector randomly, while linear initialization is to assign an orderly assignment to each neuron vector along the linear subspace of the main eigenvector of the input data.Linear initialization is used in this study. Then, in the process of neural network training, the initial weights of each neuron are continuously modified through the network competitive algorithm until the minimum Euclidean distance between the weight of the neuron vector and its initial weight is the smallest, and the change process of the neuron weight vector reflects the topological relationship between the neuron and the surrounding neurons. After hundreds of iterations (200 in this study, the setting is 200, the whole neural network tends to be stable. It should be noted that in order to eliminate the influence of sample size on SOM network training, the input layer sample data should be labeled (each value is between 0 ~ 1).

In this study, SOM classification of physical health data is based on the mapping structure of the data itself for clustering and image visualization, that is, using k-means to disposal the data. The main steps are as follows:

(1) The physical health matrix data is divided into two-dimensional image units by SOM training;

(2) The two-dimensional image units are clustered by U-matrix method, and K-means is used to cluster the two-dimensional image units to calculate the Euclidean distance of physical data visualization;

(3) U-matrix graphically depicts the relative Euclidean distance of adjacent data (gray shadow is used to show the smaller Euclidean distance, and black represents the maximum Euclidean distance, i.e., the boundary of clustering);

(4) Based on k-means algorithm, the spatial segmentation of data is carried out according to the U-matrix structure, and the visual SOM clustering graph is obtained.

In this paper, the qualitative research on the pattern recognition and visualization of the regional characteristics of students' physique is carried out by the software of MATLAB 2012.

3.2 PCA PRINCIPAL COMPONENT IDENTIFICATION AND VISUALIZATION METHOD

PCA analysis is a widely used data dimension reduction method. Its specific steps are as follows:

(1) The data standardization, the most commonly used standardization is centralization and deviation standardization. Centralization can be realized by row centralization of data matrix, column centralization of data matrix, or both;

(2) The calculation of inner product matrix between attributes;

(3) Calculation of inner product matrix. The order of eigenvalues is $\lambda \geq 1$;

(4) To find the eigenvector corresponding to the eigenvalue;

(5) The calculation the sorting coordinate matrix and calculate the information contained in each principal component (the percentage of the sum of the eigenvalues);

(6) To obtain the corresponding load of each attribute. This paper discusses the regional differences of the main influencing factors and explanatory factors of students' physical health. The visual PCA analysis is completed by R language "vegan" package, in which (1) and (2) are completed by function function RDA () in "vegan" package, (3) ~ (6) is analyzed. It is assigned by the function RDA () and the visual PCA sorting diagram is completed by the function biplot(). Finally, the physical health data are projected in PC1 and PC2 space to realize the visualization of PCA ranking chart (black spots in the plane represent all samples). The interpretation rules of PCA ranking chart are: the longer the arrow of variable is, the greater the impact of the variable on physical health (the influencing variable); the smaller the angle between the variable and PC1 and PC2 axis, the stronger the correlation between the variable and physical health. The software design is shown in Figure 3.

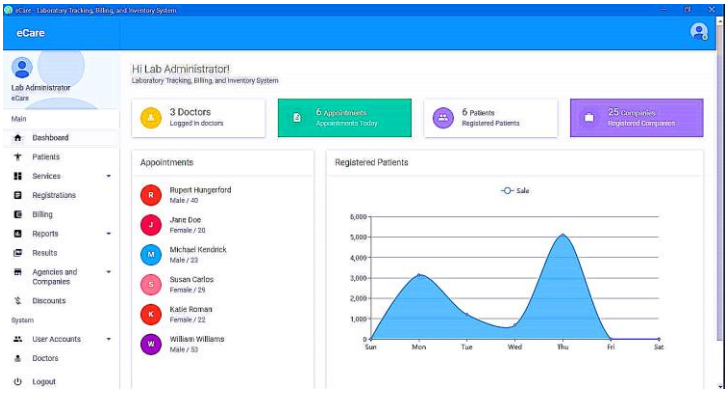

Fig. 3. The software design.

4. RESULTS AND DISCUSSION

In this study, the concept of ranking in the field of ecology was introduced to analyze the data of physical health by PCA. The visualization results in Figure 4 show the influencing factors and explanatory factors of students' physical health under different regional characteristics. In this study, PCA method will be the data set. All data through linear transformation is set to find the most representative of the original data projection method, and its projection to PC1 and PC2 in the space. Each black dot represents each single sample in the graph which are made to achieve the visual expression of PCA in the field of physical health research. At present, the research of PCA in physical health mainly focuses on the identification and extraction of main factors, data preprocessing of mathematical modeling and coupling with other models. It carried out principal component identification and extraction on the physical health indicators of firefighters in the stage of different concentric rate reserve through step test. It conducted PCA analysis on 22 indexes of cardiopulmonary function instrument coupled with ANN.

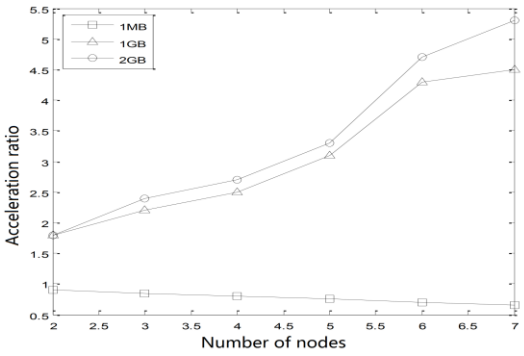

Fig. 4. The ratio test results.

5. CONCLUSION

Big data method and SOM are important methods for pattern recognition and visualization of physical health data. Taking students' physical health data of a university in Zhejiang as an example, the results of SOM and PCA analysis reveal that students' physical health has the following regional characteristics.

(1) The body weight and indicators of students' physical health have regional consistency, which is the most important factor affecting students' physical health, and also the main explanatory variable of students' physical health status in most regions;

(2) Among them, 50 m running, 1 min sit up and vital capacity constitute the explanatory variables of physical health status of regional girls; the regional differences of boys' physical health are relatively small, in addition to indicators, boys' physical health is greatly affected by body introduction and lung capacity, and some areas are also related to 50 m running and 1 000 m running;

(3) The results of visual PCA also revealed that the clustering characteristics of students' physical health

indicators had regional consistency, that is, height and vital capacity belong to the same group, and weight belongs to the same group, 50 m running and 800 m running (female) / 1 000 m running (male) belong to the same group, standing long jump and 1 min sit up (girl) / pull up (boy) belong to the same group.

ACKNOWLEDGMENTS

This paper is supported by The second batch of teaching reform projects for higher vocational education in Zhejiang Province during the 14th Five Year Plan period, titled "Research on the Three level Grading Evaluation of Professional Construction with Standards, Quality, and Excellence" (Project No.: jg20240488）;National Education Science Planning Project "Misplaced Development Strategy and Governance Innovation of Private Undergraduate Universities under the Background of Classification Evaluation" (Project No.: DGA200298）

REFERENCES

[1] Hong Zhixu, Chen Hao, Cheng Liang. Data integration and decision analysis method of social governance based on big data. Journal of Tsinghua University (NATURAL SCIENCE EDITION), 2017 (3): 1264-269.

[2] Cheng Lin, Zhu Xiaofeng, Lu Jingyun. Research on sharing logistics information platform model based on big data. Science and Technology Management Research, 2018 (15): 234-238

[3] Xu man, Shen Jiang, Yu Haiyan. Review of data driven medical and health decision support. Industrial Engineering and Management, 2017 (1): 1-13

[4] Shang Chao Wang, Han Meng, Yang Mei. Research on the design of online learning process evaluation based on big data. Modern Education Technology, 2018,28 (10): 94-99

[5] Zhou Lei, Huang Haitao, Huang Lin, et al. Research on Guizhou transportation assistant decision analysis system based on big data technology. China Transportation Informatization, 2018 (S1): 31-35.

[6] Sun Xuan, Sun Tao. Urban visual governance decision support model and application based on big data. Journal of Public Management, 2018 (2): 120-129158-159

[7] Fan Yun. Research on students' physical health information management in big data environment. Automation and Instrumentation, 2018 (4): 58-60

[8] Yu Shengquan, Li Xiaoqing. Analysis and improvement of regional education quality based on big data. Audio Visual Education Research, 2017,38 (7): 5-12

[9] Zhong Yaping, Gu houxin, Liu Peng. Analysis on the reform of physical education hierarchical teaching driven by big data of physical health. Journal of Shandong Institute of Physical Education, 2018,34 (3): 106-111