Data Analysis of Soccer Athletes' Physical Fitness Test Based on Multi-View Clustering

Bin Jiang¹, Hua Sun², Wanjian Bai¹, Hongmei Li¹, Yong Wang¹, Hongwei Xiong³ and Ning Wang²

¹State Grid Shangdong Electric Power Company, Shandong, China
²Shandong Luneng Sports & Culture Company, State Grid Shandong Company, Shandong, China
³Shandong Luneng Software Technology Co. Ltd., Shandong, China

{ls_sunht}@163.com

Abstract. Data mining is a process that uses machine learning to extract valuable knowledge from a large number of data. The sportsman has a long period of exhaustion of human body motor function. The changes of physical function signals in each period are not obvious, there is a lack of dynamic association between the body function data, which makes the connection between the data not close. Traditional data mining methods are based on association rules mining technology [1], which excavates the most relevant data attributes in the data. Once the athletes' body function data appear more obvious faults, the correlation is weakened, which will cause the deep excavation of the disease is inaccurate. In this paper, we take athletes' different body function test data as different views of body performance, adopts a multi view clustering technology to analyse the association between test items, and further optimizes the physical fitness test index. The comprehensive evaluation of athletes based on physical data is transformed into a clustering problem, which can effectively solve the athletes' physical status evaluation.

1. Introduction
For competitive sports, the ultimate goals of sports training are to enhance excellent sports performance, and athlete's physical ability is the most basic factor to improve athletic ability. Physical fitness testing is a basic way for coaches to understand the physical status of athletes [1][2]. Coaches should regularly perform physical fitness tests for athletes, according to different test standards, calculate the performance of each athlete's physical fitness test, and then evaluate their physical condition based on their own experience, and make corresponding training plan and guide training accordingly [3]. However, with the accumulation of test data [1], it is becoming more and more difficult to manage and analyse these data by manual processing. By using the traditional data processing and database management functions of computers, we can solve the management problems of physical test data, help coaches to manage athletes, translate performance scores, and manage historical data, so as to improve the efficiency of data processing [4]. However, the scientific evaluation and prediction of sports physical status cannot be solved, and the knowledge hidden behind the data resources cannot be found.

Analysis of data is local or the surface characteristics of the traditional data analysis and processing method, which cannot be predicted information about the overall features of the data hidden in the
data and then describe the trend of its development, and the physical condition of the evaluation and prediction of concern is precisely the important information hidden behind the data. Data mining technology can extract valuable, unknown, hidden and potentially useful knowledge from a large number of original data [5]. Therefore, it is very suitable for the analysis of athletes' physical test data.

The existing main methods usually use classification and association rules [6][7] in data mining to analyse athletes' physical fitness test data, on the one hand, try to find out the relationship between physical fitness test items, and to optimize test indexes. On the other hand, the classification prediction model is set up, which is used to evaluate and predict the physical status of athletes. However, with the increasing number of athletes' physical fitness test data, the dimension of data is also increasing. At the same time, the classification method belongs to supervised learning, which requires a lot of artificial annotation data, so the traditional data processing method is difficult to meet the needs of practical applications. Therefore, this paper adopts a multi view clustering method to mine the importance degree of each physical fitness test index, and cluster and evaluate athletes' physical state.

2. Existing Problems
Taking football as an example, the physical test of athletes generally includes 15 indexes, such as aerobic, anaerobic, jumping, and lower limb joints. By analysing the physical test data, the coach could find out the gap between each athlete and formulate the training program. However, with the accumulation of physical test data, the traditional processing and analysis methods of data cannot meet the actual needs, the following two points are particularly important.

2.1. Optimization of Physical Energy Test Index
At present, the 15 indicators of physical fitness in football are the coaches based on many years' practical experience, mainly involving jumping, Li, physiological and biochemical, lower extremity joints and so on. Due to the limitations of equipment, time, climate and other factors, it is very difficult to carry out the test of all 15 projects regularly. Moreover, practice shows that there is a correlation between some physical fitness items in the 15 test items, that is, the improvement of physical fitness test results can lead to the improvement of one or more physical fitness test results. For the above reasons, the coaches put forward the further optimization of physical fitness test indicators. So, what standards are optimized and optimized? First of all, we should make clear the relationship between the various physical testing projects, and provide a scientific basis for further optimization of physical energy indicators.

2.2. Assessment of Physical Status
The purpose of physical training is to improve athletes' physical quality through training, so that athletes can achieve the best physical fitness during competition and play the best competitive performance. The traditional method is highly relying on the experience of the coaches, to adjust and strengthen the best physical fitness of the athletes during the preparation of the competition. This requires the coaches to have rich experience in training and fully understand the physical state of the athletes. Obviously, this increases the pressure of the coach's decision making and the possibility of making a mistake. Now, the physical fitness test indicators have a quantitative standard, can we analyse the physical state of the athletes by these data and evaluate their physical status. And a long time test has accumulated rich test data. These data are important historical data for the physical status changes of each athlete. Through these historical data, we can analyse athlete's physical state development and change rule, predict athlete's physical state in a certain period, and provide scientific basis for coaches to make training plan.

3. The Proposed Method
During the training process, the coach can evaluate the physical status of each athlete by careful analysis of the athletes' physical test data. Taking football as an example, the coach usually divides the physical status of athletes into three states, which are poor, general and good according to the test data.
This process of evaluation is a process that coaches judge according to experience, and it is difficult to describe exactly. The clustering analysis method in data mining through data mining model potential from the characteristics of the known training data, through each data clustering to different clusters have an effective description or model, and then use this model to the unknown new data division. Therefore, the coach's comprehensive evaluation of the athletes can be regarded as a potential clustering model. In the actual life, it is necessary to evaluate the real physical fitness of the athletes by testing many different physical indexes. In order to fully test the athlete's physical state, need to consider a number of fitness test index, in this paper, a different view of the data we will a fitness test index as the physical condition of reaction of athletes, by using a method of multi view clustering [8] of athlete's physical state clustering and evaluation. The algorithm is proposed by Cai [9], which can make use of multiple physical test indicators and learn the importance of different physical test indicators through weights.

3.1. Multi-view item clustering

Let $X^{(v)} \in R^{M \times d^{(v)}}$ denote the $v$-th test index data matrix of physical fitness data, $G^{(v)} \in R^{W \times K}$ would be the clustering indicator matrix of the $v$-th test index data, $F^{(v)} \in R^{d^{(v)} \times K}$ could be the centroid matrix for the $v$-th test index data, $K$ could be the number of clusters.

As many different test index data shows that the physical state of the same athlete, thus the clustering indicator matrix $G^{(v)}$ of different test index data should be unique. This multi-view clustering method belongs to hard clustering methods, so the clustering indicator matrix $G \in R^{M \times K}$ should satisfy the 1-of-K coding mode.

The RMKMC algorithm used the structured sparsity-inducing norm, $L_{2,1}$ -norm, which could be robust to data outliers and achieve stable results with different initializations. Using those techniques, the loss function of RMKMC method can be represented as:

$$
\min_{F^{(v)}, G^{(v)}, \alpha^{(v)}} \sum_{v=1}^{V} (\alpha^{(v)})^T \| X^{(v)} - G(F^{(v)})^T \|_{2,1}
$$

$$
s.t. G_{ik} \in \{0,1\}, \sum_{k=1}^{K} G_{ik} = 1, \sum_{v=1}^{V} \alpha^{(v)} = 1
$$

where $\alpha^{(v)}$ is the weight factor for the $v$-th test index data and $\gamma$ is the parameter which controls the distribution of weights for different test indexes data, $V$ is the number of test indexes data matrix.

In order to effectively solve this object function, Cai [9] proposed an efficient optimization algorithm, which rewrite the formula (1) as:

$$
J = \min_{F^{(v)}, D^{(v)}, G, \alpha^{(v)}} \sum_{v=1}^{V} (\alpha^{(v)})^T H^{(v)}
$$

Where

$$
H^{(v)} = \text{Tr}\{ (X^{(v)} - G(F^{(v)})^T)^T D^{(v)}(X^{(v)} - G(F^{(v)})^T) \}
$$

Through this algorithm, we could alternatively update $F^{(v)}$, $G$, $D^{(v)}$, as well as $\alpha^{(v)}$, and repeat the process iteratively until the proposed objective function is convergence. For more detail of optimizations about RMKMC algorithm can be seen in [9]. Using the unique cluster indicator matrix $G$, the physical test data of each athletes could be clustered into K groups.

4. Experiments

In this section, we will validate our motivation, which applies the RMKMC method to the real soccer athletes' physical fitness test data. Our data is collected by the Shandong Luneng football club in the
physical test of the athletes. Because of the privacy of physical data, most of the data are encrypted or processed by Hash.

### 4.1. Experimental Setup

In this subsection, we compare the results of our approach with some baseline methods, includes simple multi-view K-means clustering (NKMC), affinity and Propagation (AP). We also neglect the weight of each type of physical fitness test data, and we degenerate the RMCMC method into a simple version, called simple MKMC (SMKMC), which means each test index data having same importance. We compare RMKMC approach with this simple version method. Before doing the multi-view clustering algorithm, each physical fitness test index data should be normalized, we making all the values of each physical fitness test index data in the range \([-1; 1]\). When we implement SMKMC method, we simply concatenate all of the normalized data matrix as input for the classic K-means method.

In addition, RMKMC has a parameter $\gamma$ to control the weight distribution between all physical fitness test index of data. We search this parameter as described in [9]. We use the K-fold cross validation to select the best number of clustering, and we set the number of clustering $K=5$. Since the performance of all the clustering algorithms are highly depending on the initializations, we repeat 50 times using random initialization for all the clustering methods and show the average results.

### 4.2. Clustering Results Comparisons

In this subsection, we compared different clustering results on soccer athlete data set. In this experiments, we use the NMI, ACC and Purity as our metrics to measure the clustering result of different methods, those metrics are commonly used in most of clustering methods. As shown in Table 1, the RMKMC method obtain the best clustering results.

| Method   | NMI     | ACC     | Purity  |
|----------|---------|---------|---------|
| K-means  | 0.6612  | 0.5621  | 0.6432  |
| AP       | 0.6536  | 0.5955  | 0.6422  |
| Co-Reg   | 0.7151  | 0.6632  | 0.7258  |
| SMKMC    | 0.7633  | 0.6939  | 0.7812  |
| RMKMC    | 0.7901  | 0.7798  | 0.8602  |

Compared with the SMKMC method, which just concatenate all of the normalized data matrix as input for the classic K-means method, the RMKMC method have the excellent clustering performance. This results shows the RMKMC method can make use of multiple physical test indicators and learn the importance of different physical test indicators through weight.

Figure 1. The calculated average clustering accuracy confusion matrix for Soccer Athletes’ data sets.
We plot the confusion matrices of RMKMC in terms of clustering accuracy in Figure 1. From both tables and figures, we can see that our proposed methods consistently beat the base line method on the data set.

Through the analysis of the clustering results of the athletes' physical fitness test data, we communicate with the coaches and the medical staff. The results show that the mean VO2max for 56.2ml/kg/min, the maximum heart rate 194.3/min, 4 min after exercise blood lactic acid 12.9mmol/L, anaerobic threshold heart rate 173/min, immediately after exercise heart rate 189.7/min, 3 min after exercise heart rate 130.1/min, anaerobic threshold the speed of 4.01m/s. The development of aerobic capacity should refer to the anaerobic threshold heart rate or anaerobic threshold, and the best training results can be achieved by training at different intensities.

5. Conclusions
The physical fitness test is a basic way to understand the physical condition of athletes for the coach. The traditional data mining method is based on association rules mining technology, which excavates the most relevant data attributes in the data. However, once the body appeared obvious fault, the relevance between data is weakened, will result in inaccurate association mining. In this paper, we regard different physical fitness test data as different views of athletes' body performance, and use a highly robust multi view clustering method to cluster and evaluate athletes' physical state. Finally, the experimental comparisons show the robustness and effectiveness of the algorithm.

References
[1] Pelegrini A, Silva D A S, Petroski E L, et al. Health-related physical fitness in brazilian schoolchildren: data from the Brazil sport program.[J]. Revista Brasileira De Medicina Do Esporte, 2011,17(2) :92-96.
[2] Kamangar F. Heterogeneous image feature integration via multi-modal spectral clustering[C]// Computer Vision and Pattern Recognition. IEEE, 2011:1977-1984.
[3] Yu H, Zheng D, Zhao B Y, et al. Understanding user behavior in large-scale video-on-demand systems[J]. ACM SIGOPS Operating Systems Review, 2006, 40(4):333-344.
[4] Russo S G, Neumann P, Reinhardt S, et al. Impact of physical fitness and biometric data on the quality of external chest compression: a randomised, crossover trial.[J]. Bmc Emergency Medicine, 2011,11(1):1-9.
[5] Piatetsky-Shapiro G, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data[M]. ACM, 1996.
[6] Rani B, Kaur P, Bansal A. Analysis of student physical fitness data using data mining algorithm[J]. International Journal of Applied Engineering Research, 2012, 7(11):1659-1662.
[7] Cui G X, Liang L I, Wang K K, et al. Research and improvement on Apriori algorithm of association rule mining[J]. Journal of Computer Applications, 2010, 30(11):2952-2955.
[8] Yan D, Huang L, Jordan M I. Fast approximate spectral clustering[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009:907-916.
[9] Cai X, Nie F, Huang H. Multi-view K-means clustering on big data[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2013:2598-2604.