Acta Technica *62* No. 3A/2017, 107–114

System research on pressure source's prediction and analysis for athletes with improved hierarchical K-Means algorithm

Huichao Li[1]

Abstract. With the development of competitive sports, psychological factors become the key to success or failure of athletes. The feasibility of predicting and analyzing the pressure source of athletes based on improved hierarchical K-Means algorithm is explored. Firstly, hierarchical clustering is used to obtain the results of initial pressure source clustering. Secondly, the K-Means algorithm is used to continue clustering. The improved algorithm has high speed and efficiency as well as good clustering effect. Meanwhile, this algorithm is viable to predict and analyze athlete's pressure force.

Key words. Data mining, hierarchical clustering algorithm, athletes, pressure source, predictive analysis.

1. Introduction

With the development of competitive sports, the scale and number of spectators are gradually expanded, and the intensity of competition on the sports field is also increasing. These will have some psychological impact on the athletes' psychological pressure. Stress response has become a serious problem in today's competitive sports. If athletes want to do best in stadium, they have to overcome psychological factor. Therefore, many scholars and experts pay more attention to the research field of competition pressure.

Generally, the traditional psychological counseling is usually conducted through the inquiry of psychological professionals, or some questionnaires are provided to evaluate the psychological status of athletes. However, this method has poor efficiency and cannot give effective suggestions. Through participating in the development of "Athletes Race Stress Management System" from scientific research project of the General Administration of Sport, it is known that this traditional method is

[1]Civil Aviation University of China, Tianjin, 300300, China; E-mail: huichaolicivil@163.com

applied to the competitive pressure guidance in sports field. The method of data mining is applied to cope with psychological pressure. It cannot only solve the situation of large number of athletes and complicated psychological conditions, but also effectively solve the problem of coping with stress [1].

Cluster is an unsupervised method of machine learning. It aggregates data items, observation value or feature vectors into groups. At present, clustering technology has experienced rapid development. As an important branch technology of data mining, clustering technology has a variety of applications in various fields such as machine learning, data mining, information retrieval, picture segmentation and pattern classification. It even involves biology, psychiatry, archaeology and other fields. Cluster, as one of the steps of exploratory data analysis, always has been researched and discussed by many experts in many fields. This reveals the facts that cluster have a broad appeal and practicality [2]. But, there is not a general concept and method to evaluate the cluster because of the differences between different disciplines. Clustering analysis is particularly suited to explore the internal relations within the data set and the structure evaluation due to its feature. Stream data mining is a more active research direction in data mining field. Recently, clustering algorithm is widely applied in streaming data research. Generally, streaming data is a data generated by a lot of dynamic such as network data stream [3]. Some scholars put forward an improved K-Means algorithm called - Stream KM++ algorithm. This algorithm achieves convection data mining algorithm through improving data structure and distance formula, and it also obtains a good result [4].

Therefore, clustering algorithm and content-based recommendation algorithm are applied to competitive stress psychology analysis and recommendation. After clustering, the corresponding coping strategies of the athletes are obtained, which can improve the mental state of the athletes when facing the pressure. The athletes have the ability to resist compression, and the psychological level is improved when they suffer setbacks.

2. Improved K-Means algorithm

For sports competition pressure source data, it has a small data size and large data feature. If only use the K-Means algorithm, it is found that there exist many problems in early experiment and no good discrimination. Therefore, in the cause analysis, the small amount of data and more data dimensions are considered, which may affect the clustering results. According to the characteristics of the competitive pressure source data, the improved algorithm is considered [5].

In order to obtain a better initial center and time complexity, an improved hierarchical K-Means algorithm is proposed for the traditional hierarchical K-Means algorithm. It is assumed that X is $[x_1, x_2, \cdots, x_n]$ and it is a data in n R-dimensional spaces. In view of the need to determine the K value problem ahead of time, the algorithm firstly uses a silhouette coefficient to confirm the number of cluster [6]. When reaching this level after using hierarchical clustering, the number of clusters and the initial center of the iteration are locally adjusted. In this way, it greatly saves the computation for multi-layered cluster. In addition, the criterion of intra class

similarity is adopted when the initial centers are adjusted locally. It decomposes the smallest cluster into two new clusters. In this way, a partial adjustment is carried out for the cluster with poor condensation and misclassification. This will make the choice of the initial center more reasonable and also facilitate the operation [7].

Step 1: The original data is processed and the contour coefficient is calculated by formula (1). This value is used as the initial value when the K value is the maximum

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{1}$$

Step 2: The agglomerative hierarchical clustering algorithm is used to merge adjacent two clusters to form new clusters.

Step 3: The new cluster centers of the merged clusters are calculated, and the average values of the two cluster hearts in the upper layer are calculated.

Step 4: Step two and three are repeated until the number of cluster reaches $(K - R)$ $(0 \leq R \leq K - 2)$. If $K = 2$, then $R = 0$.

Step 5: The intra cluster similarity of all the assigned clusters is calculated.

Step 6: The cluster within smallest cluster similarity is selected, that is the cluster within largest class-radius. The cluster is divided, while the cluster center c_i and the farthest sample point x_{i1} are found. In the cluster, the farthest sample point x_{i2} from x_{i1} is selected.

Step 7: Points x_{i1}, x_{i2} and other cluster center are selected as the new cluster center to do the K-Means clustering again.

Step 8: If the center of mass changes, then it returns to step 6; otherwise the algorithm ends and outputs the result.

It is observed that hierarchical clustering algorithm is used to cluster the original data first from step one to four. And then the process enters the K-Means cluster from step five to six. The number of clusters is reselected according to the number of clusters calculated by the previous hierarchical clustering algorithm. In addition, the initial clustering center of K-Means algorithm is also selected. From step 7 and 8, it enters K-Means algorithm to perform twice clustering.

The improvement of this algorithm:

1) Because the small sample data has a small size, it will male big error if we directly do K-Means algorithm. It is bad for us to analyze and decide data. Therefore, cluster is used to combine the hierarchical clustering and $-$Means.

2) The initial clustering center of K-Means algorithm no longer does random selection, and it will confirm the operation situation of former hierarchical clustering. Divide the cluster with lowest cluster similarity to save the operation time.

In order to verify the effectiveness of the algorithm, Iris Data, Breast Cancer Data and Abalone data from UCI database are selected. The data set number and characteristic number are shown in the following table. The experiment test is carried out in a PC computer (2.4 GHZ Intel CPU. 2 G Memory, Windows system).

In order to compare the algorithm performance, the K-means algorithm and the improved $H - K$ algorithm are used to cluster the data downloaded from the UCI website. The results of clustering are compared from the aspects of efficiency and

aggregation. The run time contrast time of the CPU is shown in Fig. 1.

Table 1. List of test data set

DSN Data set name	Date set size	Cluster number
Iris	150	3
Breast cancer Wisconsin	286	2
Abalone data	4177	29

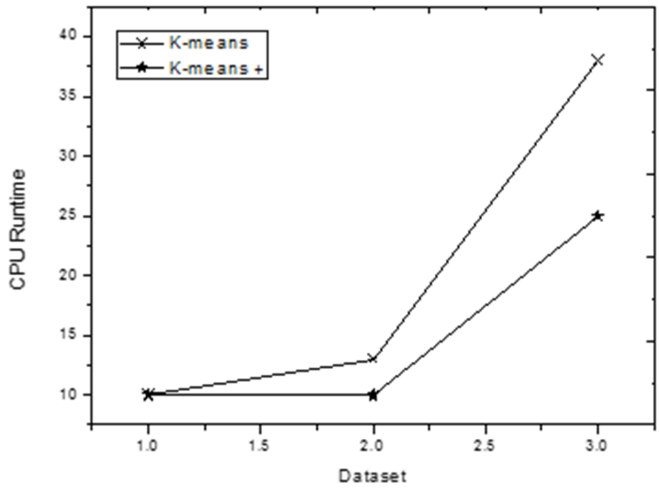

Fig. 1. CPU run time contrast broken line graph

From the above figure, as the number of data sets increases, the running time of CPU is significantly increased. the improved hierarchical K-Means clustering algorithm has a more development on run time compared to traditional K-Means algorithm, and the increasing range is obvious. Because the improved algorithm estimates the value of K in advance by using silhouette coefficient, it only optimizes the small range near the K value, which effectively reduces the time complexity of the algorithm.

In addition, in order to show the cluster degree, the accuracy rate is used to evaluate the algorithm efficiency, while iris data is used to verify cluster effect. The details are shown in Table 2.

Table 2. Accuracy comparison of two algorithm in Iris data set (%) run time

	Cluster 1	Cluster 2	Cluster 3
K-Means	100	94	72
Improved H-K	100	96	76

From above, improved $H-K$ algorithm has a high accuracy compared to traditional K-Means algorithm. Although it has a little increase, the effect is closer. It

shows that the improved algorithm has great development on efficiency and accuracy to small sample data set.

3. Application of improved K-Means algorithm in sports competition pressure source data

According to improved algorithm, it is known that cluster analysis is needed after data cleaning and transformation. Firstly, silhouette coefficient is used to work out the general cluster number K. Figure 2 shows the contour coefficient (re) value curve when K values range from 2 to 100. It is known that silhouette coefficient obtains the maximum value when cluster number reaches 2. It means that the number of clusters is better at 2, but this is not the final result. Based on hierarchical clustering algorithm, the cluster is carried out first. Then the next initial clustering center of K-Means is figured out according to the improved algorithm, which is convenient for later calculations.

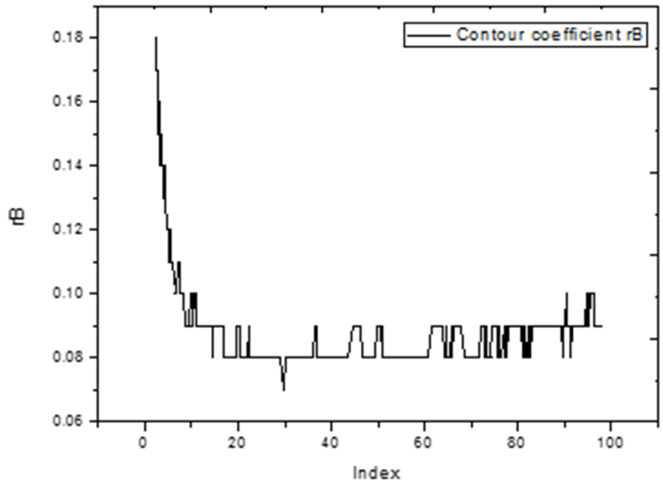

Fig. 2. Silhouette coefficient graph

The improved hierarchical K-Means algorithm is adopted to cluster the data set, clustering the 22 dimension such as competition pressure source, social support and athlete burnout. The final cluster center is shown in Fig. 3.

The overall similarity between the first classes (Series 1) and the second (Series 2) is relatively large, and parts of the property are different. The third class (Series 3) accounts for less than the total number, but it is significantly different from the first and second classes. It means this kind of athletes have a poor psychological level (low overall score shows that the pressure source is non-conformance and they have less pressure).

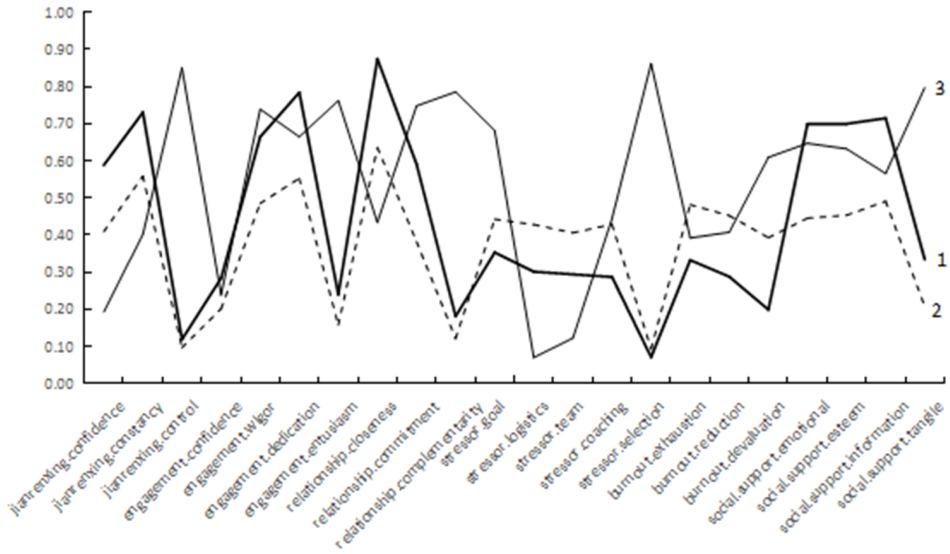

Fig. 3. Sketch map of final clustering center

4. Conclusion

In this paper, the improved K-Means algorithm will be applied to the data of sports competition pressure source. Firstly, the experiment determines the cluster number of initial hierarchical cluster according to the contour coefficient and makes first hierarchical clustering. Secondly, the initial clustering center of new K-Means algorithm is figured out again according to the improved algorithm. Finally, the K-Means algorithm is used to cluster the data sets, and the clustering results are obtained. In a word, improved algorithm has high speed and efficiency as well as good clustering effect. The improved hierarchical K-Means algorithm is feasible for predicting the pressure source of the athlete.

References

[1] Y. Xiong, J. Yang, Y. Li: *Price and carbon emission decisions under pressures of consumer, regulator and competition.* International Journal of Manufacturing Technology and Management *30* (2016), Nos. 1–2, 87–115.

[2] Y. Ding, R. Ma: *Hidden Markov model based time-series images clustering algorithm and its application in sports image processing.* Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering) *9* (2016), No. 1, 44–52.

[3] L. Zhang, F. Lu, A. Liu, P. Guo, C. Liu: *Application of k-means clustering algorithm for classification of NBA guards.* International Journal of Science and Engineering Applications *5* (2016), No. 1, 1–6.

[4] D. M. McMillan, D. J. Irschick: *Experimental test of predation and competition pressures on the green anole (Anolis carolinensis) in varying structural habitats.* Journal of Herpetology *44* (2010), No. 2, 272–278.

[5] N. N. Astakhova, L. A. Demidova, E. V. Nikulchev: *Forecasting method for grouped time series with the use of K-means algorithm.* Applied Mathematical Sciences *9* (2015), No. 97, 4813–4830.

[6] R. G. Negri, W. B. da Silva, T. S. G. Mendes: *K-means algorithm based on stochastic distances for polarimetric synthetic aperture radar image classification.* Journal of Applied Remote Sensing *10* (2016), No. 4, paper 040501.

[7] J. D. Rose: *An efficient association rule based hierarchical algorithm for text clustering.* International Journal of Advanced Engineering Technology *7* (2016), No. 1, 751–753.

Received May 7, 2017