Implementation Big Data Analysis on Football Competitions via K-means Based Tactical and Generalized Linear Model

Xin Xiang[1, a, *], Guowen Qi[2, b]

[1] *Department of sports training, Jilin sports institute, Changchun, China*
[2] *Department of sports training, Jilin sports institute, Changchun, China*
[a, *] *correspondence author email: xiang.xin.edu@outlook.com, [b] 1016103072@qq.com*

Keywords: K-means algorithm; tactical; performance indicator

Abstract: Big data analysis has been applied for technical and tactical performance evaluation in football industry and is prevalent in a decade. It is apparent that novel techniques such as mobile, e-commerce and big data are integrated in this area to inspire new innovations. This paper assesses the technical and tactical performance indicators and the competition results of 2017 China Football Association Super League Tournament via K-means cluster analysis. Competition scores are divided into two categories, balanced and non-balanced scores. A generalized linear model is built for each technical and tactical performance indicators. The competition results in score-balanced matches are defined by linear model to decipher the correlations between the performance indicators and the win probability of the competition. Furthermore, magnitude-based inferences is adopted to define the significance of the linear relationship between tactical performance indicator and the win chance of the game. The developed big data based on K-means algorithm with linear model is available to evaluate the football match performance and facilitate to design task-oriented training plan.

1. Introduction

Internet network with electronic commerce populated worldwide while it followed with rapid development of big data. Football industry has also begun to exploit big data to analyze technical and tactical performance in this decade. The analysis refers to the objective record on tactical behaviors and events in football training and competition. The process can be defined as an application of specific data to reflect training and competition tactics in many aspects of football industry area, in which quantitative analysis and cognitive strategy are served as research method to investigate multiple characteristics in football activities and competition events[1]. Based on the definition of big data application in football industry, two characters, data and quantitative relationship, are considered to be the main issue during the analysis of football tactics performance[2, 3]. However, classic investigation practically focused on on-the-spot statistical methods and video observation. Technical and tactical indicators were collected manually for statistics analysis in training and competition performance, which the data varies in different studies

[4-6]. Besides, the common features during the studies are descriptive statistical indicators, such as scoring rate, error rate, average and standard deviation[1, 4]. Such indicators are independent and non-figurative and merely displayed relatively small amount of information. It is not beneficial for in-depth judgment of training difficulties and hardly fulfills the purpose of the training[7, 8].

Generalized Linear Model (GLM) is an effective complex mathematical model which has used to determine the correlation between tactical performance-related variables and competition result difference in group sports [9, 10]. In contrast, current investigation on football technique and tactics still remains at the level of descriptive and comparative study. However, the complex mathematical models that used to define the variables are as well regarded to technical and tactical indicators and competition result variables, for instance, wins, scores, promotion, to tackle with the limitation on proper explanation of big data for evaluation of practical performance in football matches, we established an appropriate mathematical model to define the causal relationship between the value of football tactics and results of the football match. Therefore, K-means Cluster Analysis with generalized linear model and Magnitude-based Inferences were applied in the work to establish mathematical model to survey the technical and tactical performance indicators and game results in all score-balanced matches for all participating teams in 2017 China Football Association Super League Tournament. The performance was investigated in order to support reliable big data analysis on football training and matches.

2. Material and Method

2.1 Samples and Variables

The sample during the study consists of 200 games and 400 sets of technical and tactical statistical data of the 2017 season China Super League. The research variables include the results of each team's game in each match (win, flat, negative), the playing field (home and away), and the technical and tactical indicators data of each team's matches. On the basis of knowledge acquisition, 19 technical and tactical indicators were chosen. All the indicators were divided into 3 groups: Goal shot related variables (Shooting, ejection and deviation), offensive organization-related variables (control ball rate, pass, success rate of pass, passing, passing the success rate long pass, direct plug, fouling, offside, corner, header success rate) and defense-related variables (steals, steal success rate, foul, yellow cards, red cards).

2.2 K-means Algorithm

The goal of K-means is to divide n sample points into k clusters so that each point belongs to a cluster that corresponds to its nearest centroid, which is used as a clustering criterion. The centroid is the mean of all sample points within a cluster.

2.2.1 Algorithm Description

The description of K-means algorithm is as shown below:

1) Select k points randomly from the data and set as the initial centroid. Assign each point to the nearest centroid, so to form k clusters
2) Recalculate the centroid of each cluster (that is, the mean of the interior points of the class). Reassign each point to the nearest centroid, so to form k clusters until the centroid is no longer fluctuates

The K-means algorithm uses cosine similarity, Euclidean distance, or other criteria to distance metrics. The centroid is the mean of all sample points within a cluster; the number of iterations of

K-means algorithm will increase when the random effect is not ideal[11, 12].

2.2.2 Algorithm Analysis

Algorithm analysis is as shown below:

1) Processing of outliers: Outliers are generally regarded as noise which might affect the discovery of clusters and then lead to irrational experimental results. Therefore, it is essential to determine outliers before K-means.
2) Selection of initial centroids: The random selection of initial centroids may cause the case of excessive concentration, which results in an increase in the number of iterations. In this case, K-means ++ can be used to solve the issue. The K-means ++ algorithm steps are as follows in Table 1:

Table 1. K-means ++ algorithm steps

1. Given a set of initial points *D*
2. Randomly select a point from points set *D* as the initial center point
3. Calculating the distance S_i from each point to the nearest center
4. Summing S_i to get *Sum*
5. Getting random values *Random* (0<*Random*<*Sum*)
6. The set of loop points *D*, which is to do *Random* $=S_i$ operation until *Random* <0, then point *i* is the next center point
7. Looping 3-6 steps until all *k* center points are removed
8. Performing K-means algorithm

Another method is as well applicable: randomly select the first point and then take the centroid of all points as initial dataset. Along with each subsequent initial centroid, the point that is farthest from the initial centroid would be selected. The point is random and extend. However, this method may reach outliers. In addition, it is not inexpensive to find such point that is far away from the initial centroid.

Algorithm termination conditions: The function of target that reaches its optimum or the maximum number of iterations can be terminated. For different distance metrics, the objective function is often different. When Euclidean distance is used, the objective function is generally to minimize the sum of the squares of the distance from the object to its centroid (*c*), which is as shown in Equation 1.

$$\min \sum_{i=1}^{k} \sum_{x \in c_i} dist\,(c_i, x)^2 \tag{1}$$

When we use cosine similarity, the objective function is generally to maximize cosine similarity sum of the object to its cluster centroid, which is as shown in Equation 2:

$$\max \sum_{i=1}^{k} \sum_{x \in c_i} \cos ine\,(c_i, x) \tag{2}$$

K-value determination: The K-means clustering number (K-value) is defined by the user. the distribution of the data set is not known because of the initial value. K-means does not automatically learn to cluster into K clusters like EM algorithm does[13]. In order to solve this problem, K-means can be combined with hierarchical clustering. First, hierarchical clustering algorithm is used to roughly determine the number of clusters to find the initial cluster. Then K-

means is used to optimize the clustering results.

2.3 Generalized Linear Model

Let the dependent variable Y_1, Y_2, $\cdots Y_n$ be n independent observations, which is subject to exponential distribution, that is, it has a density function, which is as shown in Equation 3.

$$f(Y_i \mid \theta_i,\Phi)=\exp(Y_i\theta_i/\ \Phi\text{-}b(\theta_i)/\ \Phi +c(Y_i,\ \Phi)) \quad (3)$$

Where θ_i and Φ are parameters, b(.) and c(.) are functions. Assuming X_1 X_2, $\cdots X_n$ is the observation value of the P-dimensional argument X that corresponds to Y_1, Y_2, $\cdots Y_n$. For marked as $\eta_i = x_i^T\beta$, where β is the P*1 unknown parameter vector. Assuming $E(Y_i)=\mu_i$, and μ_i with η_i have a relationship as $\eta_i = x_i^T\beta$ $\eta_i = f(\mu_i)$, the model thus defined is a generalized linear model, θ_i is called a natural parameter, Φ is a discrete parameter and $f(.)$ is a connection function (continuous function)

One of the issues that should also be considered during mathematical models establishment is the nature of the competition, that is, the competition with balanced scores (close to the score) and the competition with unbalanced scores (A big score of wins or lose). In an unbalanced scored match, the winner is most likely to be better than the other side in all the competition tactical performance data. Both sides of the game are not available to have the best performance because of the disparity in scores and the loss of win or lose suspense. Therefore, it is vital to define the two different types of matches. Analysis of the competition with balanced scores can represent the general characteristics of the football match performance[9, 10].

2.4 Data Statistics

K-means cluster analysis is used to define the games with both balanced scores and unbalanced scores. The results showed that the game with more than 2 goals in the goal difference was an unbalanced game (goal difference (GD) 3-6 balls, (3.51±0.70) balls, total 35 games), and GD was 0-2 balls as balanced matches ((0.88±0.72) balls, total 205 games). The 390 sets of technical and tactical data of 180 balanced scores were imported into the next analysis.

The value of the technical and tactical performance indicators for the matches should be standardized. Goal shot related variables and offense-related variables are standardized to the value of our team's 50% ball control, which is as shown in Equation 4.

$$P_{\,stand\ value} = (P_{\,original\ value}/V_{\,own\ team}) * 50\% \quad (4)$$

Where, P is the value of a certain variable, and V is the ball possession of own team. The defensive-related variables are normalized to the value of the opponent team's 50% ball control, which is as shown in Equation 5.

$$P_{\,stand\ value} = (P_{\,original\ value}/V_{\,opponent\ team}) * 50\% \quad (5)$$

Where, P is the value of a certain variable, and V opponent team is the ball possession of opponent team. The 5 percentage variables of the ball possession rate, pass success rate, passing success rate, header ball success rate, and steal success rate do not require the above conversion.

A generalized linear model was created for each tactical performance indicator value and competition result in each score-balanced competition to define the linear relationship between the competition tactical performance indicators and the winning probability of the competition. In the generalized linear model, the following model is created, which is as shown in Equation 6.

$$\ln(od) = a + bx + c \quad (6)$$

Where, $od = V/1 - V$, V is win probability of the team, x is a certain technical and tactical

performance index value, a、b、c are constants. According to this model, the following derivation can be made:

Suppose that when $x=x_0$,

$$V_0=50\%,\ od_0=V/1-V=1,\ \ln(od_0)=a+bx_0+c=0.$$

$$\text{When } x_1=x_0+\Delta x,\ \ln(od_1)=a+b(x_0+\Delta x)+c=b\Delta x$$

$$\text{When } x_2=x_0-\Delta x,\ \ln(od_2)=a+b(x_0-\Delta x)+c=-b\Delta x.$$

The above two formulas are added, which is as shown in Equation 7.

$$\ln(od_1)+\ln(od_2)=\ln(od_1\times od_2)=b\Delta x-b\Delta x=0,$$

$$\text{Therefore } od_1\times od_2=1 \quad (7)$$

The above two formulas are subtracted and can be shown in Equation 8.

$$[V_1/1-V_1]\times[V_2/1-V_2]=1$$

$$OR=ratio\ of\ od=od_1/od_2$$

$$V_2\text{-}V_1=\sqrt{OR}-1/\sqrt{OR}+1 \quad (8)$$

When $\Delta x=0.5$, V_2-V_1 represents the change in the win probability of a certain football team when a certain tactical performance indicator value x changes by one unit. When $\Delta x=x$ is the standard deviation (SD), V_2-V_1 represents a certain football team's tactical performance index value, which is from a typical small value (-SD) to a typical large value (+SD), that is, when the 2 standard deviations are increased, the team's winning probability changes.

The method of magnitude-based Inferences is adopted to define the saliency of the linear relationship between each technical and tactical performance index and the probability of winning during the competition. We will substitute the two times of standard deviations (2SDs) of each tactical performance indicator value into the formula derived from step 3, it is useful to calculate the increase in 2SDs and the change (increase or decrease) within the win chance and able to calculate the 90% confidence interval of the probability change value. The saliency of win probability change value is judged by the "Smallest Worthwhile Change". In a football game, a 10% win probability change value is defined as the "Smallest Worthwhile Change". When the value of a certain tactical performance indicator increases 2 standard deviations, and the resulting change in win probability varies. The positive and negative “Smallest Worthwhile Change (±10%)” is included, the relationships of its index and win probability change is defined as significant. The two standard deviations of the three variables for the yellow card, red card, home and away are “1”, that is SDs=1, which can be calculated for each yellow card, red card, and the change value of winning probability of home and away matches.

The K-means clustering analysis and the generalized linear model creation are completed by the data statistics software SPSS 20.0. The magnitude-based inference is calculated by Excel 2010.

3. Results and Discussion

During the 180 balanced competitions of the 2017 season China Super League, the original data and standardized data of the technical and tactical performance indicators for averaging per football team per competition are shown in Table 2.

The linear relationship between the technical and tactical variables and the competition results in the 2017 season China Super League balanced scores deduced from the generalized linear model developed during the study and showed in Figure 1. It revealed that the number of shots (winning

probability changing value; ±90% confidence interval: 15.8±13.9), positive number of shots (35.4±17.1), number of passes (20.3±16.1), success rate of pass (26.7±16.4), number of straight plugs (17.1±22.4) and number of steals (14.3±13.8) can significantly increase the winning probability of the football team. There is a significant negative correlation between the increasing in the number of fouls (-24.9±17.4) and red cards (-31.0±25.8) with the probability of winning the ball. Increased header success rate (-7.5±16.9), steal success rate (-7.1±15.0), number of foul (-4.3±13.7) and yellow card (0.4±5.1) cause a slight difference to the team's winning probability. The competition situation variables of home and away can bring 9.5% (± 90% confidence interval: ± 14.9) increments for the competition winning probability; other variables have no significant correlation with the winning probability of the team.

Table2. Descriptive statistics data of technical and tactical performance indicators for each competition (±s)

Variable	Original value	Standard value
Shooting/N	13.05±4.98	13.05±4.35
Ejection/N	4.56±2.65	4.67±2.65
Deviation/N	8.05±3.82	8.01±3.14
Control ball rate/%	49.95±9.35	
Pass/N	380.79±90.21	380.51±52.38
Success rate of pass/%	78.54±6.65	
Passing/N	19.54±7.98	19.21±6.12
Success rate of passing/%	25.32±10.87	
Long pass/N	55.04±13.24	56.89±17.05
Direct plug/N	0.75±1.54	0.78±1.56
Fouling/N	15.84±4.76	16.35±5.69
Offside/N	2.03±1.85	2.14±1.95
Corner/N	4.91±2.78	4.83±2.45
Header success rate/%	50.03±12.78	
steals/N	15.65±4.95	15.98±5.55
steal success rate/%	81.14±10.87	
foul/N	17.04±4.57	17.65±5.84
yellow cards/P	1.94±1.11	1.96±1.21
red cards/P	0.10±0.31	0.09±0.29

The linear relationship is expressed to add 2 standard deviations for the tactical variable of a certain game，which is the change value in the win probability. Based on Black dots in Figure 1, positive value indicates increasing, and negative value indicates decreasing. The error line in Figure 2 is the 90% confidence interval for the change value, and the vertical dashed line shows the Smallest Worthwhile Change (±10%). During the study, the aim of the K-Means algorithm is to

determine K cluster centers randomly and classified the sample points to each cluster according to the nearest neighbor principle. The K-Means algorithm in the study was applied to divide all competitions of the 2015 China Football Association Super League into balanced and unbalanced matches. The competition with balanced scores uses a generalized linear model to effectively define the causal relationship in each football competition between the technical and tactical performance indicators and the winning or losing of the games, so that it is a valuable method to determine which competition-oriented technical and tactical indicators are the winning indicators. The information above provided by the model is reliable practically for competition performance assessment, opponent information detection and training plans design.

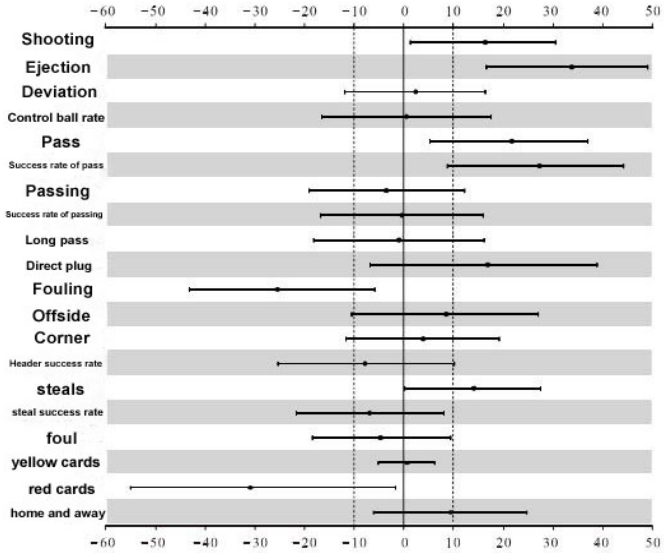

Figure 1. The linear relationship between the technical and tactical variables from various football competitions and the competition results during the 2017 China Super League matches with balanced scores

References

[1] Z. Yue, H. Broich, J. Mester, Statistical analysis for the soccer matches of the first Bundesliga, International Journal of Sports Science & Coaching 9(3) (2014) 553-560.

[2] C. Carling, A. Williams, T. Reilly, The Handbook of Soccer Match Analysis. Abingdon, UK: Routledge (2005).

[3] C. Carling, A. Williams, T. Reilly, The handbook of soccer match analysis: A systematic approach to performance enhancement, London: Routledge, 2005.

[4] H. Liu, M.-Á. Gomez, C. Lago-Peñas, et al., Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup, J. Sports Sci. 33(12) (2015) 1205-1213.

[5] H. Liu, Q. Yi, J.-V. Giménez, et al., Performance profiles of football teams in the UEFA Champions League considering situational efficiency, International Journal of Performance Analysis in Sport 15(1) (2015) 371-390.

[6] A. M. Williams, Perceptual skill in soccer: Implications for talent identification and development, J. Sports Sci. 18(9) (2000) 737-750.

[7] M. D. Hughes, R. M. Bartlett, The use of performance indicators in performance analysis, J. Sports Sci. 20(10) (2002) 739-754.

[8] R. Mackenzie, C. Cushion, Performance analysis in football: A critical review and implications for future research, J. Sports Sci. 31(6) (2013) 639-676.

[9] K.-Y. Liang, S. L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73(1) (1986) 13-22.

[10] J. A. Nelder, R. J. Baker, Generalized linear models, Wiley Online Library1972.

[11] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data mining and knowledge discovery 2(3) (1998) 283-304.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, et al., An efficient k-means clustering algorithm: Analysis and

implementation, IEEE transactions on pattern analysis and machine intelligence 24(7) (2002) 881-892.
[13] A. K. Jain, Data clustering: 50 years beyond K-means, Pattern recognition letters 31(8) (2010) 651-666.