2014

ISSN : 0974 - 7435 Volume 10 Issue 14

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 10(14), 2014 [7617-7622]

The application of PCA and cluster analysis to the NBA match evaluate

Ronghui Pan
Sports Department of Tianjin Polytechnic University, Tianjin, 300387, (CHINA)

ABSTRACT

PCA is helpful to analysis the factors influencing the NBA match results. Cluster analysis can also, through classifying 30 NBA teams based on their own features and referring to the datum of the recent three seasons, make a scientific analysis of the match influences. Both PCA and cluster analysis can scientifically study factors effecting the NBA match, providing necessary datum for the team reform and adjustment. So PCA and cluster analysis can meet the new requirements that the traditional way of analyzing the team datum cannot meet, laying a solid foundation for the further study.

KEYWORDS

PCA; Cluster analysis; the NBA match evaluate; Applied research.

INTRODUCTION

PCA's datum exploration and its analyzing aspect enable it to clearly show the close connection between datum. Cluster analysis can make the differences between datum more distinct by classifying those datum, making it possible for us to do a detail analysis. The necessary conditions for our NBA match evaluate created by PCA make it more easier for us to find influencing factors, offering strong support to the team's further reform and improvement and providing favorable conditions for improving team's performance. This paper introduces the efficient evaluating of the NBA match through PCA and cluster analysis, both of which enable the evaluating process scientific and correct.

THE INTRODUCTION OF CLUSTER ANALYSIS THEORY

What is cluster?

Cluster means the status's classifying process, in which the datum of every class share much commons with each other. However there are many differences existing in different classes. Through deeply exploring the datum, the potential connections between the different classes can be completely reflected, enabling the data structure to show the corresponding character.

The similarity measure

The definition of the cluster analysis above has shown us that the datum in the same class share high similarity, so the specific character of data can be fully reflected, which is shown clearly in Figure 1 bellow.

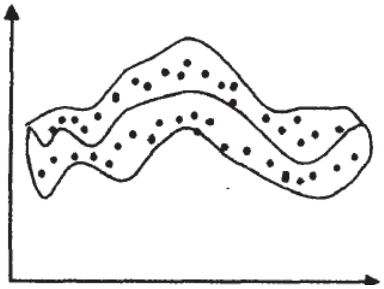

Figure 1: The strip class

Classifying the splashes in figure 1 through different means can achieve different classes, fully showing the similarity of the different datum. This is also the main principle of the data classifying process and from another aspect reflects the positive influence of the similarity on the cluster analysis. As a result, the similarity measuring algorithm comes on the scene.

The efficient classification of the datum through cluster analysis in the data exploring process can clearly reveal the necessary connection between different data, making the data classifying more scientific and providing a necessary support to the data density measure, which makes it possible to build a data analyzing model. By this way there can form a connecting link between different datum. The cluster datum show different shapes in the exploring process, which plays a promoting role in classifying datum. The shape is shown in Figure 2 bellow:

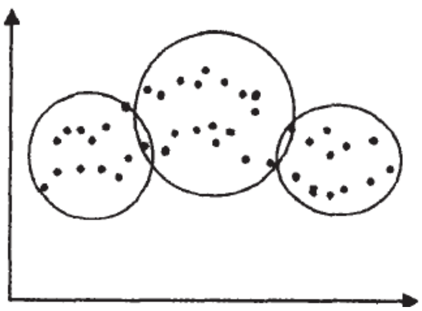

Figure 2 : The globular class

Through the efficient analysis of the different NBA teams and with a goal of using the PCA, this paper makes an efficient luster analyzing to datum, clearly reflecting the class of different datum. Besides, this paper also researches the data similarity, providing a valuable evaluation for the sustainable development of the NBA match.

Data processing

Statistical method and principle

Getting the statistics and analyzing datum through CPA should involve in the multivariate statistic based on both quantitative and variable values, making it easier for researchers to find specific components and enabling their original information to be saved completely, which here generally refers to the quantitative information, not to other parts.

The original data processing

We use statistic analysis software to process the principal components of the original data, and in this process the characteristic value of measuring factors chosen by us is all above 0.5, which enables us to do similarity research for these measuring factors.

The result and the analysis

Original data

We have efficiently collected the datum of the NBA regular and post seasons of 2010-2011,2011-2012,2012-2013, enabling the original datum more persuasive and real.

Principal Component Analysis

Choosing ζ_1 and ζ_2 as two data indexes, then choosing series samples from ζ for researching the season datum. ζ_1 and ζ_2 should be included in η_1、$\eta_{2,}$which are within two ovals in the long and short axis. The index factors we research should share connections with the indexes of ζ_1 and ζ_2 and enable $\eta_1 \perp \eta_2$。 to be involved in the principal component processing, in which η_1 acts as the projected coordinate of ζ （ζ_1， ζ_2） in the long axis, and η_2 as the projected coordination in the short axis. Through building the projection coordinate, we can observe that N observed value creates corresponding changes, which are caused by the volatility of η_1 in the long axis. However the volatility of η_2 in the short axis. Through the above analysis, we can see that η_1 can be studied as the comprehensive index and η_2 can be treated as an auxiliary indicator. So the changes of N observed values can be reflected more clearly. In this paper, η_1 is also treated as the principal component of the season data researching. In conclusion, to build a projected axis in the principal component analysis can efficiently reflect the influence of the low dimension space on the high dimension space, remaining the original characters and nature of datum. In the season data researching, we have made an efficient collection of the relative datum through the official website, which includes 10 regular season datum and 7 datum with representative characters. After collating these datum, we choose the datum of 16 teams taking part in the post season and research them through the principal component analysis. This process is shown in TABLE 1 below, through the datum in which we can get the analysis results of several seasons datum, which are similar to ones achieved through the analyzing theory.

TABLE 1 : The total variance explained of 2012-2013 season datum

component	original characteristic value			extract and load quadratic sum		
	total vale	Variance %	accumulated value %	total vale	Variance %	accumulated value %
1	2.345	33.504	33.504	2.345	33.504	33.504
2	1.761	25.156	58.659	1.761	25.156	58.659
3	1.120	16.002	74.661	1.120	16.002	74.661
4	0.827	11.810	86.471	0.827	11.810	86.471
5	0.432	6.165	92.636			
6	0.342	4.880	97.516			
7	0.174	2.484	100.0			

In the total variance explained list of 2012-2013 season data, the principal component of 7 datum has been respectively analyzed, leading to such a conclusion that the characteristic values of 4 datum among total in list 1 are all above 0.5 and their accumulating contribution rates are all above 86%, which is much similar with the result of analyzing other two season's datum. So we can get that among seven characteristic factors, four of which play an important role in team's achievements in every season, and that they are shooting average, three-point shooting average, free shooting average and rebounding control average.

ANALYZING NBA COMPREHENSIVE ABILITY THROUGH CLUSTER ANALYSIS

According to the NBA team's achievements in recent three seasons, we collect statistics of ever team's free throws, three-point shooting average, two-point shooting average and the amount of rebounds in every match, all of which have been analyzed through PCA in list 1 and we also get the similar result with this. Except them, there are some other indicators, which can also to some extent influence team's achievements. They are error rate, assist rate, the foul average and the steal average. In the process of collocating other datum, we would input them in a professional statistic analyzing software to study their influence on team's performance through cluster analysis. The analyzing process shows us that in recent three seasons, 30 teams can be classified into three classes, every team in which share less differences in influencing factors but more similarities in them. Based on cluster analysis, we classify Pacers, Miami Heat, Raptors, Chicago Bulls, Washington Wizards, Spurs, Oklahoma Thunder, Clippers, Houston Rockets and Trail Blazers as the first class; Nets, Charlotte Bobcats, Hawks, K nicks, Cavaliers, Warriors, Grizzlies, Mavericks, Suns and Timber wolves to the second class; the last teams of Nuggets, Pelicans, Kings, the Lakers, Utah Jazz, Detroit Pistons, Boston Celtics, the Philadelphia 76ers, Bucks and the Magic to the third class. The seventeen datum of each class have clearly reflected that ever team has their own characters. the 17 datum's average values are shown in TABLE 2. The main difference indicators of three classes can be analyzed through TABLE 1.

TABLE 2 : The main indexes of every class

event/class	the first class	the second class	the third class
two-point hitting	20.85	18.26	16.43
two-point shooting	38.14	36.08	38.87
2 points rate	0.55	0.51	0.43
three-point hitting	9.30	8.40	6.43
three-point shooting	23.22	24.04	20.23
three points rate	0.41	0.35	0.32
the free-throw shooting	16.04	13.73	13.87
the free-throw hitting	22.78	19.02	20.23
the free-throw rate	0.71	0.72	0.68
offensive rebound	9.59	9.20	9.67
defensive rebound	25.69	22.08	20.43
rebound	35.28	31.27	30.10
assisting attack	16.87	14.02	9.80
foul	19.69	21.05	17.90
steal	8.10	6.44	7.17
fault	11.63	13.06	13.93
block shot	2.88	2.13	2.60

In terms of two-point hitting, free-throw shooting, defensive rebound and assisting attack, the datum of teams in the first class are obviously higher than those of teams in other two classes. This enable us to make a conclusion that the feathers of excellent teams are having accurate two-point hitting, badly damaging the opponent team and enjoying more free throws, focusing on defending box card and stealing, having more assistants, acting as a whole, players cooperating with each other and highly demanding the arrangement of the defender position.

The features of teams in the second class reflecting in datum are that the datum about two-point hitting, free throws, rebound and assisting attack are lower than those of teams in the first class. These teams have less two-point shooting, more three-point shooting and faults. So their features can be concluded that they focus on periphery attack and defend actively and fiercely.

The datum of teams in the third class are completely lower than those of ones in the first and second class. Their perimeter shooting is less accurate than teams in the first class and their defending is less fierce than teams in the second class.

THE APPLICATION OF CLUSTER ANALYSIS THROUGH FUNCTION

The similarity coefficient

In cluster analysis, Pearson's similarity coefficient can be used to measure the differences in curves through function, and the formulation is below:

$$\rho = \frac{\int_o^T \left[x_1(t) - \frac{1}{T}\int_0^T x_1(t)dt \right]\left[x_2(t) - \frac{1}{T}\int_0^T x_2(t)dt \right]}{\sqrt{\int_0^T \left[x_1(t) - \frac{1}{T}\int_0^T x_1(t)dt \right]^2 dt \int_0^T \left[x_2(t) - \frac{1}{T}\int_0^T x_2(t)dt \right]^2 dt}}$$

In scientific calculating of the similarity coefficient, the standardizing of the function curve can highlight curve's morphological features, enabling the similarity coefficient to replace the specific distant between separating points of analysis objects in cluster analysis. Besides these coefficients should be reclassified according to the relative principles, sharing commons with the traditional classifying analysis.

Realizing data standardization through function

The analyzing process of data standardization introduced in this paper is not the traditional standardization process of eliminating dimension's effects but the one targeting to diminish and finally eliminate the real distant of function items, making their average value reach 0, variance 1, both of which enable functional datum to be efficiently clustered after being standardized. The formulation reflecting this process is below:

$$x_{xd} = \frac{x(t) - \frac{1}{T}\int_0^T x(t)dt}{\frac{1}{T}\int_o^T \left[x(t) - \frac{1}{T}\int_0^T x(t)dt \right]^2 dt}$$

Referring to the data standardization method, datum can be clustered through the function formulation. There are three main steps: firstly, calculating the range of function data within the required time, based on which getting vectors of the similarity coefficient; secondly, calculating the vectors of corresponding coefficients based on the functional data; finally, doing cluster analysis on the basis of the results of two steps above.

Through the introduce above we can see that in analyzing of function data standardization, cluster datum share some commons with each other; the first one is the collocation of data, in the process of which datum can be proceed better. Besides, there also exists some differences between them, such as the changing of cluster's distance. But clustering the different aspects simply through the function datum cannot distinctly reflect the difference between function datum and the real distant.

The clustering of the function data

When analyzing the clustering process of the function data, we can see that there are much similarity existing in results and many commons in their characters, both of which are results hardly achieved through traditional cluster analysis. The traditional clustering analysis focuses on the real distance between datum. But because the distance between two classes is short, so the classification is not clear. And we should pay more attention on rules and connections of datum. Because not allowing the reversing process, the function cluster analysis is more scientific and persuasive. Derivative method's dynamic reflection of the function curve's features enable function datum to cluster derivatives, making the differences in function shapes be clearly reflected.

In data cluster analysis, the cluster should be calculated based on datum introduced in this paper and function datum's characters. This is finished through two steps: firstly, clustering the similarity datum of function distant after classifying function datum and their cluster. In this step function datum's distant can be efficiently calculated through clustering functions, and the similarity of datum in the same class can be calculated with the help of function data.

Firstly, we should calculate the distance in the process of defining cluster. The function data $x_i(t) = 1,2,...,n$ can be multiplied within $\left[t_1, t_2\right]$ and $x'_1(\ t\ ) = 1,2\ldots n$ is its derived function.

The distance of function datum of $x_1\left(t\right)$, $x_2\left(t\right)$ within $\left[t_1, t_2\right]$ is

$$d(x_1, x_2) = \left(\int_{t_1}^{t_2} (x_1(t) - x_2(t))^2 \right)^{1/2}$$

The distance of derived function of $x_1^{'}\left(t\right)$, $x_2^{'}\left(t\right)$ within$\left[t_1, t_2\right]$ is

$$d(x_1, x_2) = \left(\int_{t_1}^{t_2} (x'_1(t) - x'_2(t))^2 \right)^{1/2}$$

The mean function of function data $x_1\left(t\right) i = 1,2\ldots n$ is

$$\frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

Through the way of producing function datum above, we can fit the record datum as function ones, which enable us to cluster these function datum.

CONCLUSION

Through analyzing function data, we analyze the clustering of NBA season's datum. And factors influencing NBA team's achievements have also been efficiently studied through systematically analyzing the indicators in every class by cluster analysis, enabling us to get a correct evaluation of team's performance in match. So this method can play a positive role in promoting team's development and providing strong theory and data support for analyzing team's achievements.

REFERENCES

[1] Lina, Zhou Weibo, Wang Jinfeng; The building and application of ANFIS model based on PCA, Yellow River, **4**, **36(6)**, 80-83 **(2014)**.

[2] Huang Liwen; The nearly ideal analyzing method of CPA and its application. Mathematical statistics and management, **32(6)**, 1013-1019 **(2013)**.

[3] Ren Zhiqiang, Tan Hongxiang, Pan Wenliang; Using PCA to evaluate cigarette product's quality and stability, Tobacco science & technology, **2**, 5-8 **(2013)**.

[4] Zhang Zhenwei, Zhang Nana, Shilei; The exploring of Chinese Medicine Preparation evaluation model based on PCA, Chinese journal of experimental formulas of Chinese medicine,**19(16)**,18-21 **(2013)**.

[5] Yang Xiaoxia, Tian Shengkui; The analysis of Chongqing tourism's developing potential based on PCA. Journal of Southwestern University: nature and science edition, **4**, 111-117 **(2013)**.

[6] Liu Genxia; The ranking of listed company's comprehensive competition based on PCA--taking electronic industry as example, Communication of Finance and Accounting: comprehensive (Chinese), **12**, 15-17 **(2012)**.

[7] Zhang Yingdong; The application of advanced PCA to the multiple-responding optimization, Modular machine tool and its automated processing technology, **11**, 97-100 **(2012)**.