Analysis of NBA Team Strength Using Players' Race Data Based on Clustering Method

Mingzhe Xu and Junhui Gao

***Abstract*—In this paper, we use k-means, a basic type of clustering analysis, to analyze the data downloaded from the official site of NBA. We analyze the relevance between the number of clusters a team has and its winning percentage, the correlation coefficient is not high. However, we find that when the assuming number of clusters is ten, the relevance between the research items is the highest.**

***Index Terms*—NBA, clustering analysis, correlation, statistics, winning percentage.**

I. Introduction

Data analysis has already been a popular trend of this modern society. We have entered an so-called era of big data and data analysis has permeated into many kinds of fields, including on the basketball court.

Early in 2006, a research report had shown that some basketball coaches began to use shot charts to reflect players' preference to make shots to help improving the team and also make plans to play against their opponents. [1]

Muthu Alagappan, a fourth-year Biomechanical Engineering student at Stanford University and a consultant at Ayasdi, is crazy about sports and the statistics they include. [2] In 2011, he collected huge amounts of data from the official site of NBA and put them into a simulation program of Ayasdi, a company he works for, and the results he got shocked the whole world. According to the scatter diagram, he categorized NBA players into thirteen positions, which conclusion was totally different from the traditional one. This research made huge impact at that time because it redefines the positions on the court and solve the problem of oversimplification.

In 2016, Zhengyang Xie from China, he chose 25 star players of all five positions from NBA official website, downloaded their data and analyzed the data from many aspects. He found out that all the chosen players prefer to make shots right in front of the basketry. In addition, there are two climaxes at the location of the basketry and the three-point line. In the aspect of time, it shows that the attempts of two-point shots reach the lowest and those of three-point reach the highest during the last five minutes. [3]

In this paper, we use the method of clustering to process the players' data on the court in a whole specific season. In the first part, we review some previous researches about data analysis in NBA. We introduce both the method of clustering analysis and our own method fits our analysis. In the second section, we do simulation and calculation, including comparisons between the winning percentage and the number of clusters of each chosen team. In the third part, we discuss different possibilities of the number of clusters we simulated and try to find which number of clusters best test to analyze a team. The last part is our conclusion of this research. We concluded that when the assuming number of clusters is ten, the relevance between the research items is the highest. In other words, ten clusters of players are the fittest when used in this research in the second part.

Manuscript received August 17, 2017; December 12, 2017.

Mingzhe Xu is with Wuxi Big Bridge Academy, Jiangsu, China.

Junhui Gao is with the American and European International Study Center, Wuxi, China (Corresponding author: Junhui Gao; e-mail: jhgao68@163.com).

II. Data and Method

A. Data Collection

Our data comes from NBAstat [4], we choose the data of all the players who played in 2014-2015 and 2015-2016 regular seasons.

The statistics include teams, average starting time, average time, average field goal percentage, average rebound, average free throw percentage, average assist, average turnover, average block, average steal, average foul and average point per game.

As for a few blank items in the original item, we replace them with zero.

B. Clustering Analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis.

Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.

Clustering algorithms can be categorized based on their cluster models. [5]

C. Our Method

The data for the regular season of 2015-2016 includes a total of 475 players and 30 teams. However, some players do not belong to any teams, so we put them into a single team. 475 is divided by 31, each team has an average of 15.3225 players.

We use the method of clustering analysis to cluster all the players into several groups. Then we count how many clusters each team has and compare the number of clusters with the

doi:10.18178/ijssh.2017.7.12.919

winning rate of the regular season, trying to find the relevance in it.

III. CALCULATION

A. Use k-means to Cluster all the Players

K-means clustering is a method of vector quantization, originally from signal processing, that into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

1) Let C1 . . . Ck be the initial cluster centers.
2) For each point di in D, assign it to the closest cluster Cj such that violate-constraints (di, Cj) is false. If no such cluster exists, fail (return {}).
3) For each cluster Ci, update its center by averaging all of the points dj that have been assigned to it.
4) Iterate between (2) and (3) until convergence.
5) Return {C1 . . . Ck}.

K-means clustering is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows:

1) Each instance di is assigned to its closest cluster center.
2) Each cluster center Cj is updated to be the mean of its constituent instances. The algorithm converges when there is no further change in assignment of instances to clusters. [6]

B. Analysis of Relevance in Teams' Winning Percentage

When we use k-means, we have to appoint the number of clusters, and we get the corresponding number for each team through investigation. At the same time, we list the winning percentage (regular season) of each team, assuming that we cluster in the number of ten, results are shown in Table I below.

TABLE I: NUMBER OF CLUSTERS AND WINNING PERCENTAGE OF EACH TEAM

Team	Number of Clusters	Winning percentage	Team	Number of Clusters	Winning percentage	Team	Number of Clusters	Winning percentage
Warriors	9	89.0	Pacers	9	54.9	Nuggets	8	40.2
Spurs	8	81.7	Trail Blazers	8	53.7	Bucks	7	40.2
Cavaliers	10	69.5	Pistons	7	53.7	Kings	8	40.2
Raptors	8	68.3	Mavericks	7	51.2	Nicks	9	39.0
Thunder	7	67.1	Grizzlies	8	51.2	Pelicans	7	36.6
Clippers	8	64.6	Bulls	9	51.2	Timberwolves	8	35.4
Celtics	8	58.5	Wizards	8	50.0	Suns	7	28.0
Heat	8	58.5	Rockets	8	50.0	Nets	7	25.6
Hornets	8	58.5	Jazz	8	48.8	Lakers	6	20.7
Hawks	8	58.5	Magic	8	42.7	76ers	6	12.2

From Table I we can find out that teams like Lakers and 76ers have the fewest number of clusters, which are six, while cavaliers have the largest clusters in a number of 10. The number of clusters in other teams is between 7 and 9.

We analyze the relevance between the number of clusters a team has and its winning percentage, according to the statistics in Table I, we get the square of R is 0.3524. If we first square root the winning percentage, the coefficient of relevance R increases: the square of R is 0.3908, and R is 0.6251, shown in Fig. 1.

In Fig.1, the x-axis is the number of clusters a team has, and the y-axis is the square of winning rate of each team in 2014-2015 regular season.

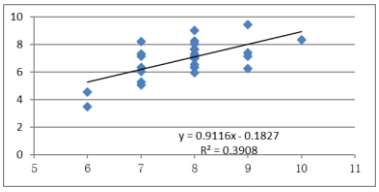

Fig. 1. Relevance between the number of clusters and winning rate.

C. Comparison between Teams

We choose three teams: Warriors with the highest winning rate in the regular season, Grizzlies in the middle, and 76ers with the lowest winning rate. The statistics are in Table II, and we will analyze these three teams one by one in the following.

76ers: It only has six types of players: only one in type 3, two in type 0, 1, 8, three in type 2, and four in type 9.

Grizzlies: It has eight clusters out of ten: one in type 0, 2, 5, 6, two in type 1, 9, and three in type 4, 8.

Warriors: Besides type 5, warriors have at least one player in each cluster. It has most players in type 4, two players each in type 0, 8, one player each in type 1, 2, 3, 6, 7, 9.

IV. DISCUSSION

In this part we consider different central number of clusters and calculate the subsequent coefficient of relevance R. When we divide the players into 8, 10, 12, 14 clusters, the coefficient of relevance between the number and winning rate

is relatively large, shown in Fig. 2.

TABLE II: COMPARISON BETWEEN THE CLUSTERS OF THREE CHOSEN TEAMS

	0	1	2	3	4	5	6	7	8	9
76ers	Robert Covington	Richaun Holmes	Jahlil Okafor	Carl Landry					Kendall Marshall	Hollis Thompson
	Isaiah Canaan	Elton Brand	Nerlens Noel						Christian Wood	Nik Stauskas
			Jerami Grant							Tony Wroten
										T.J. McConnell
Grizzlies	Matt Barnes	Brandon Wright	Zach Randolph		JaMychal Green	Mike Conley	Marc Gasol		Jordan Adams	Jordan Farmar
		Jarell Martin			Vince Carter				Elliot Williams	Tony Allen
					Xavier Munford				Russ Smith	
Warriors	Klay Thompson	Festus Ezeli	Andrew Bogut	Kevon Looney	Marreese Speights		Draymond Green	Stephen Curry	Ian Clark	Andre Iguodala
	Harrison Barnes				Leandro Barbosa				James Michael McAdoo	
					Shaun Livingston					
					Brandon Rush					

In addition, we also analyze the data of the regular season of 2014-2015, shown in Fig. 3.

Fig. 2 and Fig. 3 both show that the coefficient of relevance between number of clusters and winning percentage is not very high, especially for the regular season of 2014-2015. However, we can see that the curves for two regular seasons are similar, and when the number of clusters is ten, the coefficient is relatively large.

Besides, for any team, the distribution of clusters is uneven. For example, when a team has four players in one cluster or more, maybe it will only have one player in other clusters. This is one of the factors that we do not consider, and this is where we can improve in our further research.

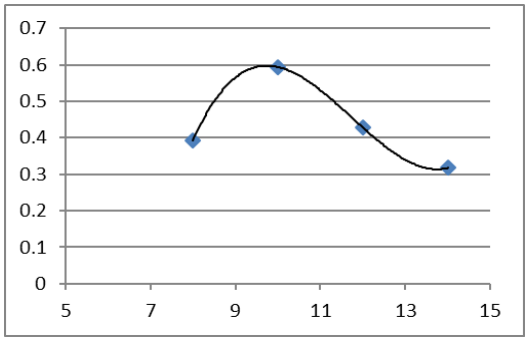

Fig. 2. Number of clusters a team v.s Winning rate (2015-2016 regular season).

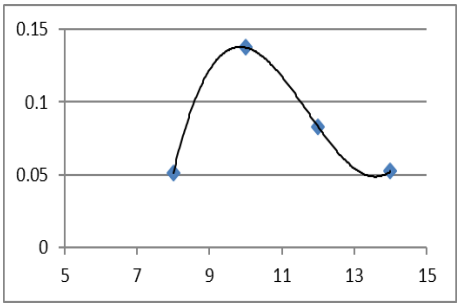

Fig. 3. Number of clusters a team has v.s Winning rate (2014-2015 regular season).

V. Conclusion

In this paper, we use k-means, a basic type of clustering analysis, to analyze the data downloaded from the official site of NBA. After using clustering analysis and making linear regression diagrams, calculating the coefficient of relevance R and discussing under different premises, we make the conclusion that the winning rate of a team is related to the number of clusters it has, which matches the conclusion made by Muthu Alagappan in his previous research. However, what is different is that ten clusters are the best and the fittest to analyze a team in our research, while he got thirteen positions in his research.

References

[1] B. J. Reich, J. S. Hodges, B. P. Carlin, and A. M. Reich, "A spatial analysis of basketball shot chart data," *The American Statistician*.
[2] M. Alagappan, "From 5 to 13 redefining the positions in basketball," presented at Sports Analytics Conference, 2012.
[3] Z. Y. Xie, "Data analysis instance for NBA star shooting," *Open Journal of Social, Sciences*, vol. 4, pp. 1-8, 2015.
[4] Stat-Nba. [Online]. Available: http://www.stat-nba.com/
[5] H. X. Ma, *Application of Cluster Analysis*, 2014.
[6] K. L. Wagstaff, C. Cardie, S. Rogers, and S. Schrodl, "Constrained K-means clustering with background knowledge," *International Conference on Machine Learning*, 2001.

Junhui Gao was born in July 1968, who is a native of Hangzhou, China. Now he is a teacher of American and European International Study Center, Wuxi, Jiangsu. He got his bachelor's degree in ecology at East China Normal University, a master's degree in computer science at Shanghai Jiao Tong University.

He is good at data mining, machine learning, mathematical modeling and software development. He has some project experience in molecular diagnosis, computer aided drug design, medical informatics, intelligent energy and other engineering fields, he has some research experience in bioinformatics, molecular simulation, computational physics and other academic fields.

He once worked at Shanghai Center for Bioinformation Technology and College of life science, Soochow University.