International Journal of Science and Engineering Applications
Volume 5 Issue 1, 2016, ISSN-2319-7560 (Online)

Application of K-Means Clustering Algorithm for Classification of NBA Guards

Libao ZHANG
Department of Computer Science, Shandong University of Science and Technology, Qingdao 266590, China

Faming LU
Department of Computer Science, Shandong University of Science and Technology, Qingdao 266590, China

An LIU
Department of Computer Science, Shandong University of Science and Technology, Qingdao 266590, China

Pingping GUO
Department of Computer Science, Shandong University of Science and Technology, Qingdao 266590, China

Cong LIU*
Department of Computer Science, Shandong University of Science and Technology, Qingdao 266590, China

Abstract: In this study, we discuss the application of K-means clustering technique on classification of NBA guards, including determination category number, classification results analysis and evaluation about result. Based on the NBA data, using rebounds, assists and points as clustering factors to K-Means clustering analysis. We implement an improved K-Means clustering analysis for classification of NBA guards. Further experimental result shows that the best sample classification number is six according to the mean square error function evaluation. Depending on K-means clustering algorithm the final classification reflects an objective and comprehensive classification, objective evaluation for NBA guards.

Keywords: K-Means clustering algorithm, NBA guards, classification number

1. INTRODUCTION

In this study, K-means clustering technique is applied to the classification and evaluation for NBA guards. Recently, the classification of NBA guards is mainly based on the starting lineup, time, points and rebounding [10]. Meanwhile, starting point guard, reserve guard, point guard and offensive guard are also frequently used in traditional classification methods.

According to traditional classification methods, researchers needed to assign classification threshold to each indicator manually, which was so subjective that some particular players could not be classified in a logic situation. In this study, K-Means clustering technique origin's from machine learning field is applied to the classification of NBA guards. In order to realize the objective and scientific classification of NBA guards, this study depends on NBA 2014-15 season guards' data which is standardized and processed by mathematical models and Java language. In this way, the guards' type could be defined scientifically and properly based on classification result. Meanwhile, the guards' function in the team could be evaluated fairly and objectively. K-means clustering and improvements is widely used in present study, such as network intrusion detection [3], image segmentation [4], and customer classification [5] and so on. A cluster analysis of NBA players are very common, but their works mainly focus on the position of players.

2. K-MEANS APPLICATION

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups. It is an important human behavior. K-means algorithm [1, 2] is the most classic division-based clustering method, is one of the ten classical data mining algorithms. The basic idea of K-means algorithm is: k point in the space as the cluster centroids to cluster, classify their closest objects [8]. Through an iterative approach, in each successive update the value of cluster centroids until get the best clustering results so that the obtained clustering satisfy objects in the same cluster have high similarity and at the same time objects in the different cluster have low similarity. Therefore, based on K-means clustering algorithm one can identify the guard's function in the team, and helps people to obtain an objective evaluation about guard's ability.

2.1 K-means models establishment

2.1.1 Data filtering and processing

The data of tables obtained from DATA-NBA (www.stat-nba.com), as shown in Table 2-1, As the main task of guard is the score, rebounds and assists, so we can select these three data items as data factors for distance calculation of clustering analysis. In addition, assists and score are different, a player 10 assists in the difficulty, not less than 20 points, if not to take the measures standard, the cluster will not be fair, the score will become the main indicator, and rebounds and assists will become a secondary indicator. So we use the following equation to deal with the data processing.

$$SP_{ij} = \frac{C * n * P_{ij}}{\sum_{i=1}^{n} P_{ij}}$$

SP_{ij} = the standard value of the player$_{ij}$

P_{ij} = the original value of the player$_{ij}$

C = auxiliary parameter for data amplification

n = the total number of players in a dataset

2.1.2 K-means algorithm defect

K-means algorithm has some drawbacks [4]: First, the number of k cluster centers need to be given in advance, but in practice the selected k value is very difficult to estimate. It is extremely difficult to know how many types of data collection should be divided in advance. Second, K-means need to artificially determine the initial cluster centers, different initial cluster centers may lead to a completely different clustering results.

Table 2-1: The part of the original data

Player	Season	Team	Rebounds	Assists	Steal	Blocks	fault	foul	Scores
Russell - Westbrook	14-15	hunderbol	7.3	8.6	2.1	0.2	4.4	2.7	28.1
James - Harden	14-15	Rockets	5.7	7	1.9	0.7	4	2.6	27.4
Stephen Curry	14-15	Warriors	4.3	7.7	2	0.2	3.1	2	23.8
Kobe Bryant	14-15	Lakers	5.7	5.6	1.3	0.2	3.7	1.9	22.3
Carey - Owen	14-15	Knight	3.2	5.2	1.5	0.3	2.5	1.9	21.7
Klein - Thompson	14-15	Warrior	3.2	2.9	1.1	0.8	1.9	1.6	21.7
Dwyane - Wade	14-15	Heat	3.5	4.8	1.2	0.3	3.4	1.7	21.5
Damian - Lillard	14-15	Blazers	4.6	6.2	1.2	0.3	2.7	2	21
DeMar - DeRozan	14-15	Raptors	4.6	3.5	1.2	0.2	2.3	2	20.1
Kevin - Martin	14-15	mberwolv	3.6	2.2	0.8	0	1.9	1.9	19.5
Chris Paul	14-15	Clippers	4.6	10.2	1.9	0.2	2.3	2.5	19.1
Isaiah - Thomas	14-15	Celtics	2.1	5.4	0.6	0	2.6	2.1	19
Monta - Ellis	14-15	Mavericks	2.4	4.1	1.9	0.3	2.5	2.5	18.9
Victor - Oladipo	14-15	Magic	4.2	4.1	1.7	0.3	2.8	2.6	17.9
Kyle - Lori	14-15	Raptors	4.7	6.8	1.6	0.2	2.5	3	17.8

Considering the first defect, we need to evaluate different values of k in the k means clustering, and select the most reasonable k value.

Considering the second defect, we choose the initial center point by the remote-first algorithm [9]. The basic idea of the initial clustering center point lies in: the initial clustering centers should be as far as possible from the distance between each other.

Detailed steps of the k clustering center with remote-first algorithm is explained as follows:

Step1: Choose one center uniformly randomly from the data points.

Step2: For each data point x, compute D(x), the distance between x and the nearest center that has already been chosen.

Step3: Choose one new data point randomly as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to D(x) 2.

Step4: Repeat Steps 2 ~ 3 until k centers have been chosen.

2.2 K-means algorithm

2.2.1 Data Preparation

In order to construct the K-means model, one needs to get the 14-15 season NBA guard data which includes 120 NBA guards' data. We standardize and filter the data, to prepare for the K-means analysis. The filtered data is stored in csv file & an excerpt of our processed data is shown in Table 2-2.

Table 2-2: 120 NBA Guard Regular Season Data

	Player	Team	Rebounds	Assists	Scores
1	Russell - Westbrook	Thunder	7.3	8.6	28.1
2	James - Harden	Rockets	5.7	7	27.4
3	Stephen Curry	Warriors	4.3	7.7	23.8
4	Kobe Bryant	Lakers	5.7	5.6	22.3
5	Carey - Owen	Cavaliers	3.2	5.2	21.7

6	Klein - Thompson	Warriors	3.2	2.9	21.7
7	Dwyane - Wade	Heat	3.5	4.8	21.5
8	Damian - Lillard	Trail Blazers	4.6	6.2	21
9	DeMar - DeRozan	Raptors	4.6	3.5	20.1
10	Kevin - Martin	Timberwolves	3.6	2.2	19.5
11	Chris Paul	Clippers	4.6	10.2	19.1
12	Isaiah - Thomas	Celtics	2.1	5.4	19
13	Monta - Ellis	Mavericks	2.4	4.1	18.9
……	.....	.....	……	……	……
……	.....	.....	……	……	……
114	Jose - Calderon	Knicks	3	4.7	17.3
115	Jason - Richardson	76ers	3.5	2	17.2
117	Quincy - Pondexter	Pelicans	3.1	1.5	17
118	Bojan - Bogdanovich	Nets	2.7	0.9	16.9
120	Marcus - Thornton	Celtics	1.9	0.9	16.6

2.2.2 Algorithm Design

Using K-means clustering algorithm for data analysis. The basic idea of K-means algorithm [11] is: allocating data set D into k clusters. To determine k clusters, we need to determine the k center C1, C2…Ck, calculate the distance to each point to the center for each point inside dataset, the point that the shortest distance from the center classified as represented by clusters.

K-means algorithm steps are explained follows:

Step1: Determine the number of K-means clustering center k;

Step2: The use of remote-first algorithm to initialize the center of k;

Step3: The points of dataset D assigned to the nearest center, forming a k clusters;

Step4: The calculation k Category cluster centroid obtained by [3], the nearest point of dataset D from the centroid as the new center;

Step5: Repeat [3] ~ [4], until the center remain stable.

Euclidean distance is calculated as follows:

$$D = \sqrt{\sum_{k=1}^{n}(P_{ik} - P_{jk})^2}$$

$D =$ *the distance between* P_i *and* P_j

$P_{ik} =$ *the value of* P_i

$P_{jk} =$ *the value of* P_j

2.2.3 K value determination

After calculation the results of the k are 2, 3, 5, 6, 7, and 8 by the k-Means algorithm, and then we use the Mean Squared Error to perform the comparison of results with different k values. The calculation formula is as follows:

$$MSE = \frac{\sum_{i=1}^{n}(P_i - PC_i)^2}{n}$$

$n =$ *the total number of point in a dataset*

$C =$ *the numbers of clustering center*

$P_i =$ *the point* i

$PC_i =$ *the center of the point* i

$MSE =$ *the mean squared error*

According to Figure 2-1 and Table 2-2, we can see that as k-values gradually increase from 2 to 8, the mean square error getting smaller and smaller. Clustering result also gradually changed for the better, and the small changes of clusters to achieve a relatively stable state when the center points surpass six. This is the minimum mean squared error, it can be concluded that when the cluster number is 6, the mean squared error is becoming smaller, the similarity within the class is higher, and classification result is the best at the same time.

Table 2-2: Mean Square Error for Different Values of Time

K-Values	2	3	4	5	6	7	8
Mean square error	6.913933	5.701685	5.023356	6.27497	4.363918	4.335483	4.323653

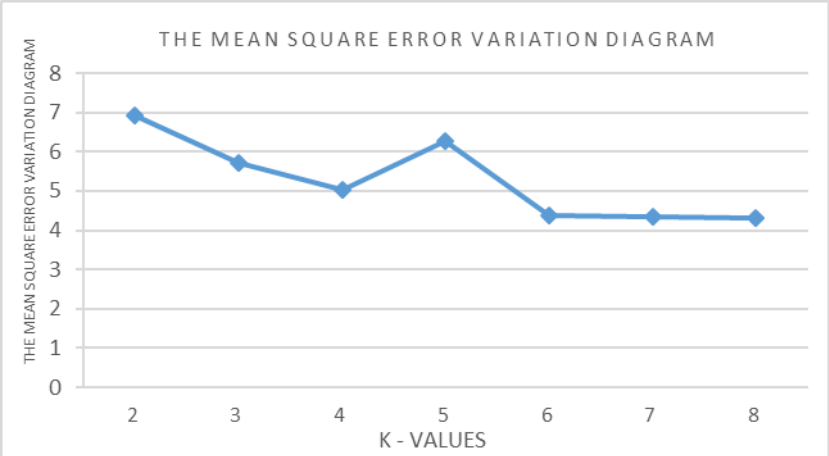

Fig2-1: The Mean Square Error Variation Diagram

3. RESULTS and EVALUATION

3.1 Classification Results

Based on the above analysis that the effect of clustering is the best when k = 6, the NBA guard can be divided into 6 categories, the classification results is shown in Table 3-1, classification and analysis of the results are as follows:

Category 1: The guards whose assist and score is well are the main shooting and point of the team. However, limited in playing time, the data is not particularly outstanding, such as Manu Ginobili, Tony Parker.

Category 2: The guards whose score ability and rebounds ability are high, with more playing time, is absolutely super guard and the core of the team, such as Harden, Curry and Westbrook.

Category 3: The guards whose score ability and rebounds ability are outstanding, assists ability is normal, are the guards of the Swingman type. They can make enough contribution to the team's defense and offense, such as Iman - Shumpert, Wesley - Matthews, etc.

Category 4: The guards who get 14.17 points and 8.27 assists, are typically assists madman, the initiator of the offense, the core and leader for a team such as Chris - Paul, John -Wall.

Category 5: The guards whose score ability are much higher than rebounds ability and assists ability, should be the team's point guard, the team's playmaker such as Wade, Owen.

Category 6: Compared rebounds ability and assists ability, score ability is the main contribution of this category guard, usually as the team's backup point guard, the outstanding ability of singles or long shot well, such as J.J-Redick, Nick - Young.

Table 3-1: K-Means Clustering Result

Classification Number	Category Centroid			Category members(Separated "|" between names)
	Rebounds	Assists	Scores	
1	2.47	4.49	11.77	Isaiah - Thomas | Brandon - Jennings | Tony - Tony Parker | Morris - Williams | Eric - Gordon | Terre - Burke | Brandon - Knight | Ismail - Smith | Jarrett - Jack | Morris - Williams | Jeremy | JR Smith | DJ- Augustin | Manu Ginobili | - Mario Chalmers | Zach - Lavin | CJ Watson | Dennis - Schroeder | Jameel-Nelson | DJ- Augustin | Gray Davis - Vasquez | Jose - Calderon |
2	5.43	6.87	20.36	Russell - Westbrook | James - Harden | - Stephen Curry | - Kobe Bryant | Damian-Lillard | Kyle - Lori | Eric - Bledsoe | Tyreke - Evans | Michael - Carter - Williams |
3	3.49	2.26	12.56	Kevin - Martin | Wesley-Matthews | Brad - Bill | Aaron - Afflalo | Avery-Bradley | Alec - Burks | Shabazz-Muhammad | A Long - Afflalo | JR Smith | Pop | Dion - Waiters | Rodney - Stuckey | Ben - Mark Lehmer | Gerald-Henderson | Jordan-Clarkson | Langston- Galloway | -Gary Neal | Will-Barton | Patrick- Beverly | Wayne-Ellington | Imran-Shumpert | Jason- Richardson | Quincy-Pondexter |
4	4.21	8.27	14.17	Chris Paul | Reggie Jackson | John Wall | Jeff - Teague | Thailand - Lawson | Zhu-Huo Ledi | Ricky - Rubio | Rajon-Rondo | Deron Williams - Williams |
5	3.68	4.78	16.61	Carey - Owen | Dwyane-Wade | Klein - Thompson | DeMar-DeRozan | Isaiah - Thomas | Monta-Ellis | Victor-Oladipo | Brandon - Knight | Derek - Ross | Kemba - Walker | Morris - Williams | Tony - Rothen | Golan - Dragic | Darren - Collison | George - Hill | - Mike Conley | Alec Frank - She Weide | Joe - Johnson | Evan - Turner |
6	2.18	1.96	11.2	J.J. Redick | Jamal - Crawford | Louis - Williams | Nick - Young | Isaiah - Buchanan | Avon - Fournier | Dion - Waiters | Aaron - Brooks | Tim - Hardaway II | OJ- Mayo | Jodi - Meeks | Anthony - Morrow | Aaron - Afflalo | AJ-Price | Alec Frank - She Weide | Courtney - Lee | - Gary Neal | Alec Frank - She Weide | Norris - Cole | Terrence - Rose | - Gary Neal | Marco – Marco Belinelli | Isaiah - Buchanan | Bojan - Bogdanovich | Marcus - Thornton |

3.2 Analysis and Evaluation

In news and media, guards are divided into point guard and shooting guard according to the arrangement in the team, and divided into key guard and reserve guard according to playing time order. Therefore, general guard has four categories: key point guard, key shooting guard, reserve point guard& reserve shooting guard. However, basketball is the athletic sports of constant adjustment and adaptation. Throughout the league process, every NBA guard assignment, as well as playing time, playing order required to make specific arrangements according to needs of the team and coach's strategy. Therefore, this intuitive classification is dependent on people's subjective judgment which is limited biased & changing. Because guards' function in the game would constantly adjustment, classification of guards should constantly adaptation, which caused a great disturbance to classification and evaluation of NBA guards macroscopically. Accordingly, the above classification and evaluation methods heavily depend on so many subjective factors, that the classification and evaluation of NBA guards are neither scientific nor objective.

In this study, the K-Means clustering analysis is applied to the classification of NBA guards. We take fully advantage of the statistical data of NBA guards to analyze data and standardize data rationally. Mining the authentic classified information, will get classification of NBA guards more scientifically and objectively. Find guards in the team's role, the ability to guards and defender in the team's performance has a comprehensive understanding and evaluation. Identify the guard's function in the team, can help people have a comprehensive understanding and objective evaluation about guard's ability and their performance has a comprehensive understanding and evaluation. Identifying the guard's function in the team could help NBA Sports News, NBA commentator and Basketball enthusiasts have a comprehensive understanding and objective evaluation about guard's ability and their performance. Furthermore，the classification results propose an effective solution for analysis the extremely big of NBA data, rather than just make statistical comparisons.

4. CONCLUSIONS

Traditionally, clustering is viewed as an unsupervised learning method for data analysis. In this study, we proposed a simple and qualitative methodology to classify NBA guards by k-means clustering algorithm and used the Euclidean distance as a measure of similarity distance. We demonstrated our research using k-Means clustering algorithm and120 NBA guards' data. This model improved some limitations, such as manual classification of traditional methods. According to the existing statistical data, we classify the NBA players to make the classification and evaluation objectively and scientifically. Experimented results show that this methodology is very effective and reasonable. Therefore, based on classification result the guards' type could be defined properly. Meanwhile, the guards' function in the team could be evaluated in a fair and objective manner.

5. REFERENCES

[1] Jiawei Han. Data Mining Concepts and Techniques [M].Beijing: Mechanical Industry Press .2006.

[2] http://en.wikipedia.org/wiki/K-means_clustering.

[3] Liu Chang Qian. K-means algorithm improvements and network intrusion detection application [J]. Computer simulation .2011.

[4] Yan Xinge .ISODATA and fuzzy K-means algorithm applied in image segmentation [C]. Chinese Optical Society 2004 Academic Conference.

[5]Qu Xiaoning .K-means clustering algorithm in commercial banking customers classification [J]. Computer simulation .2011.

[6] Raymond T. Ng and Jiawei Han, CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE TRANSACTIONS ONKNOWLEDGE and DATA ENGINEERING. 2002.

[7] Zhu Xian based on simulated annealing Particle Swarm Optimization techniques of genetic data biclustering research [M]. Nanjing Normal University .2009.

[8]Yin Z.D .Based collaborative filtering Trusted Service Selection [M]. Nanjing University of Posts and Telecommunications.2013.

[9] Jiangwen Rui. Distributed machine learning framework based on cloud [M]. Xiamen University .2013.

[10] Data Source: http: //www.stat-nba.com/.

[11] Sun Jigui, Liu Jie, Zhaolian Yu clustering algorithm [J] Journal of Software 2008.

[12] Jin Ming. Optimization Selection and Evaluation of Technical Index Classification of NBA Elite Guard of. China Sport Science and Technology. 2005.

[13] Richard J. Roiger, Michael W. Geatz, Data Mining a tutorial-based primer, Addison-Wesley, 2003.

[14] Josef Cihlar, Rasim Latifovic, Jean Beaubien. "A Comparison Of Clustering Strategies For Unsupervised Classification," Canadian Journal of Remote Sensi.