Analyzing NBA Teams Using Big Data Technologies

Supervisor: Dr. Tristan Glatard

Student: Yan Ma

Student ID: 40024856

Project URL: https://github.com/MonkeykingYan/NBA

Abstract

National Basketball Association(NBA) has entered a faster-paced offensive era. In this new era, how to manage the team structure is very importance. A team manager has to make a team with the changing era,

Section 1 is the introduction part, I give a brief description of the project background and the project objectives.

Section 2.1 give a brief description of databases, allPlayers.csv and Linups_11To16.csv. Section 2.2 is for player clustering. I select 20 features to represent a player, including Points, Rebounds, Assists, etc. Because some of the features are not independence, using PCA reducing features from 20 dimension to 3 dimension. Using K-means algorithm to clustering players from 2011 to 2016. According to "Elbow Method", select $K = 10$, where K is the number of clusters. Player clustering results are shown in section 3, Table(A), for more information, please check **playerClusters.csv** file under Clustering. The python code of this part, please check **playerCluster.py**.

Section 2.3 is for team clustering, same as section 2.2, I use K-means to classify teams. The feature of each team is constructed by team lineups, where I have a detail discussion in section 2.3.1. The distance between features are represented by Damerau–Levenshtein distance and centroids are updated according to the sum of distance to others in the same cluster. For team clustering results are shown in section3, Table (B), for more information, please check **teamrClusters.csv** file under Clustering. The python code of this part, please check **teamClassifier.py**.

Section 2.4 is for team structure discussion. In this part, I mainly discussing the players combination of teams in the same cluster, and player importance of GSW 2016 and MIA 2013. For players combination, using frequent item sets to get the most used player combination. The results are shown in table5 .The python code of this part, please check **FrequentitemSets.py**. According to the passing network between players, we can construct a team passing network, using page rank algorithm to calculate the importance of each players in a team. The results are shown in section3, table2 and 3.The python code of this part, please check **Network.py**.

Section 3 shows all the results mentioned in previous section and section 4 is the discussion part. In this section, I give a discussion on how to establish a team in one of the cluster. And using $GSW2016$ and $MIA2013$ as example analysis the different between teams in cluster 0 and teams in cluster2.

Keywords: K-means, Frequent itemsets, Page-Rank

1 Introduction

National Basketball Association(NBA) has entered a faster-paced offensive era, the league has skewed towards taking more 3-point shots due to their high efficiency as measured by points per field goal attempt. Therefore, "Small Ball"favored by more and more teams. Small ball is a style of play that sacrifices height, physical strength and low post offense/defense in favor of a lineup of smaller players for speed, agility and increased scoring. Therefore, NBA games played in 20 years ago have a big difference with today's game. In this project, I will analyze team structures in recent 5 years.

Objective: What is the structure for different teams and what kind of players are importance in this team?

2 Methods and Materials

2.1 Database

In this project, I mainly use the following 2 databases. All the data sets can be download in Stats NBA and Basketball Reference.

allPlayers.csv: The size of this database is 18206×31. This database store all the players stats from $1980-2016$. Player stats including basic information of a player, including FG, FGA, FG%, TRB, AST, STL, PTS etc.

Lineups_11To16.csv: The size of this database is 1801×20. This database store teams lineup stats from $2011-2016$. The stats including team linups, minutes played per game etc.

2.2 Player Clustering

2.2.1 Data Preprocessing and Feature Selection

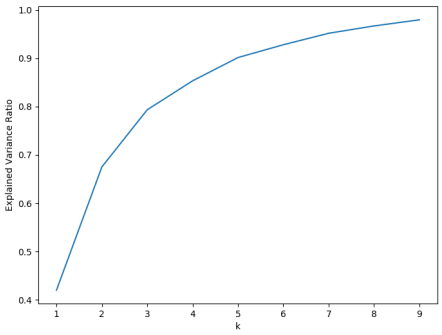
- Selecting all the players played from 2011 to 2016 where 'minutes played per game' is more than 20 minutes and 'played games' are more than 50 games. i.e. $MP > 20$ and $G > 50$.
- Rescaling the range of features to scale the range in $[0,1]$. Using formula: $data' = \frac{(data - min(data))}{(max(data) - min(data))}$ where $data$ is an original value, $data'$ is the normalized value.
- Selecting player features base on player stats. Selected features are shown in Figure1.
- Reduce dimension using PCA. Selecting $k = 3$, as shown in Figure (A), 3 dimension is able to represent almost 80% of the origin data. where explained variance ratio is: $EVR = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \lambda_{k+1} + \cdots + \lambda_n}$.

```
# All the features
FEATURES_COL = ['fg', 'fga', 'fg3', 'fg3a', 'fg2', 'fg2a', 'ft', 'fta', 'orb', 'drb', 'trb',
                'ast', 'stl', 'blk', 'tov', 'pts', 'fg_pct', 'fg2_pct', 'fg3_pct', 'efg_pct']
```

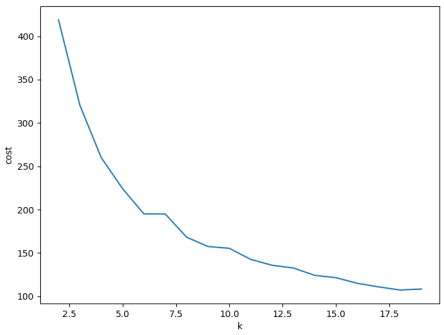
FIGURE 1: Player Features Selection

2.2.2 K-means for Playesr Clustering

For K selection, I tested K in $[1,20]$, as shown in Figure (B). According to " Elbow method", I chose $K = 10$. The results are discussed in section 3.1.



(A) Explained variance ratio



(B) K v.s Costs

2.3 Team Feature Constructing and Team Clustering

2.3.1 Team Features Construction

A team is a combination of different kind of players. The idea for team classification is lineups classification. The feature constructing algorithm is as follows:

- Step1: For each team rank the lineups by MP (minutes played per game).
- Step2: Select top3 lineups for each team. For players in selected lineup, ordering them by their position, i.e. $PG > SG > SF > PF > C$
- Step3: Constructing features as: $F = \frac{\sum_{l \epsilon Lineup} MP(l) \times l}{sum(MP(l))}$
- Step4: K-means clustering.

2.3.2 Team Classifier

Using K-means for team classification. Different with player classification, a team feature is the combination of players label. A Non-Euclid distance K-means algorithm is necessary.

Damerau–Levenshtein distance is a string metric for measuring the edit distance between two sequences. Considering a team feature is a string and use **Damerau–Levenshtein distance** to calculate the distance between each feature and update centroids with the string which has the minimum sum of distance to other strings.

For K selection, I calculated the cost for different K from $[0,15]$, shown in figure 3, when K is in range $[0,5]$, the cost decreasing sharply compared with $K > 5$, Therefore, I select $K = 5$

The resluts are shown in 3 and the overall output is under **'/Clustering/teamCluster.csv'.**

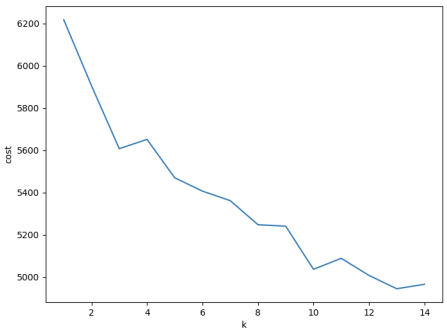


FIGURE 3: K selection for team Clustering

2.4 Team Structure Discovery

Player combinations are important for team structures. e.g. A coach may use a combination of 'break through' player with a shooter more than other combinations. I am going to use **FP-growth algorithm** to find the frequent item set for each team cluster. The 'minSupport' value equal to 0.5 and 'minConfidence' equal to 0.5.

In a basketball game, if a player ball handling time is more than others, it means this player play an importance role in this team structure. According to each teams ball passing data, we can easily construct a passing network. Using **Page Rank Algorithm** to analysis players importance in their teams. The results are shown in section 3.

3 Results

Player Clustering Result: Figure 4 illustrates that the classifier can get a clear result. The overall output is under **'/Clustering/playerClusters.csv'**. Table (A) is a small part of the output file, as shown in the table, some players are clustered into different label in different year and different team. This is reasonable as players may change their play styles with their increasing age or in different team.

Team Clustering Results: Table (B) shows a small part of the output file, for the overall results, please check **"/Clustering/teamsClusters.csv"**, the results will be discussed in section 4. For the implementation part, please check **TeamClassifier.py**

Team Construction Discovery Results: FP-growth algorithm results are shown in table5, for the implementation part, please check **FrequentItemSets.py**. Page Rank Algorithm results are shown in table 2,3 and Figure 5,6. For the implementation part, please check **network.py**

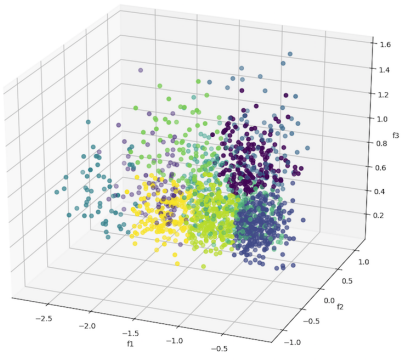


FIGURE 4: Player Clustering Visualization

Player	Team	yr	Label
Adams, Steven	OKC	2015	8
Adams, Steven	OKC	2011	0
Allen, Ray	BOS	2011	3
Allen, Ray	MIA	2011	2

(A) Players Combinations in Cluster

Team	Year	Label
MIA	2013	0
ORL	2011	1
GSW	2016	2
LAC	2011	3
DAL	2012	4

(B) Players Combinations in Cluster

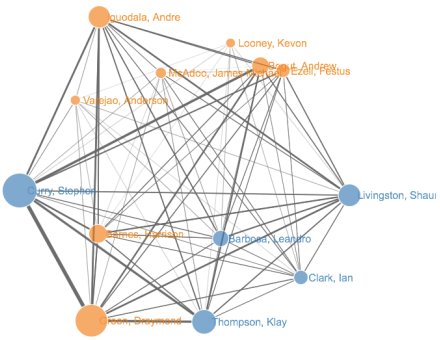


FIGURE 5: A table beside a figure

Team	Player Name	Page Rank	Label
GSW2016	S.Curryn	2.17	1
	D.Green	1.99	6
	K.Thompson	1.34	6
	S.Livingston	1.29	2
	A.Iguodala	1.21	4
	H.Barnes	0.86	5
	A.Bogut	0.77	9

TABLE 2: A table beside a figure

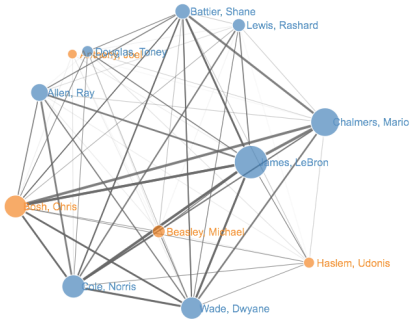


FIGURE 6: A table beside a figure

Team	Player Name	Page Rank	Label
MIA2013	L.James	1.97	1
	M.Chalmers	1.61	2
	N.Cole	1.19	2
	C.Bosh	1.157	7
	D.Wade	1.188	1
	R.Allen	0.765	2
	S.Battier	0.594	2

TABLE 3: A table beside a figure

4 Discussion

Label	Reference Explanation	Template Player
0	Common laborer	S.Adams
1	Terminator /Super Star	K.Bryant
2	3D players (3 pointer + defence)	S.Battier
3	All Round Guard	R.Allen
4	Pure Shooter	M.Miller
5	Defender	T.Allen
6	Team Organizer	C.Paul
7	Quick Big Man	S.Josh
8	All Round Big Man	K.Garnett
9	Traditional Big Man	A.Bogut

TABLE 4: Players Combinations in Cluster

Starting from player clustering results, the results are sorted by player's names. According to the clustering results, Table 4 illustrates the corresponding players with their label.

Base on the player classification results, I classify teams from 2011 – 2016 into 5 clusters. **For teams in cluster0**, according to player combination results shown in Table 5, player with label 3 play an important role in these teams. The most used combination is [0,3],[5,3] and [0,8,3], it obvious that these teams are established by a guard, at the same time, these teams may not have a good offensive big man. To contruct team cluster 0 we need, guard + defender or guard + defensive big man.

For teams in cluster1, the only 2 player combinations shown in the results. i.e. player label 7 and label 2. According to Table 4, teams in cluster1 are constructed with a good perimeter player and a good post player.

For teams in cluster2, this cluster team is very similar to cluster0, as player label 0 play an important role in these teams. However, this guard player should have better shooting skill compared with cluster 0 as they have more combination with player label 0 and 7, players 0 and 7 are used to take the rebounds after guard made the shots.

For teams in cluster3, these teams have good current depth of the lineup, as they can have a double guard combination, double big man combination and guard with big man combination. Therefore, these teams can be consider as 'Teamwork' teams.

For teams in cluster4, these teams are established with a good team organizer. To construct a team as cluster4, a shooter and a big man are necessary. It's easy to explain as team organizers' job is to get other guys into offense, including shooter and pos big man.

Label	Antecedent	Ante-Freq	Consequent	Conse-Freq	Conf
0	0	29	3	36	0.931
	5	27	3	36	0.926
	[0,8]	24	3	8	0.917
	6	22	3	36	0.909
	8	32	3	36	0.906
1	7	22	2	33	0.926
	2	33	7	22	0.7575
2	0	19	3	30	0.947
	2	19	3	30	0.947
	7	24	3	30	0.917
	6	21	3	30	0.905
	6	21	7	24	0.809
3	3	30	8	31	0.867
	6	27	7	26	0.864
	7	26	8	31	0.846
	8	31	3	30	0.839
	6	22	3	30	0.819
4	2	34	6	36	0.912
	8	29	6	36	0.897
	3	25	6	29	0.880
	[8,2]	25	6	36	0.880
	8	29	2	34	0.862

TABLE 5: Players Combinations in Cluster

Then, I took $MIA2013$ with label0 and $GSW2016$ with label2 as example, Figure 6 and Figure 5 are represent $MIA2013$ and $GSW2016$ respectly, where each node is a player, the bigger the nodes the more important they are, the link represent the passing times in one season, the link thickness means the connection between 2 players and the color represent player positions, blue one represent perimeter players and yellow one represent pos players. We can easily find 3 main difference between 2 teams.

- $GSW2016$ has more players in the passing network, which means more players took part into games. Compare with $MIA2013$, $GSW2016$ has more balance pagerank value, which means a player will have more oppotunity to participate in the game.
- $MIA2013$ the ball are always in perimeter players hands, as we analysis before, team with label 0 start from perimeter players.