Classification of NBA players

Zibo Song & Tao Wang

1 Introduction

Inspired by the topological data analysis of Muthu Alagappan in "From 5 to 13: Redefining the Positions in Basketball" and the "Periodic Table of NBA Elements" by Stephen Shea in Basketball Analytics: Spatial Tracking, this study proposes an alternative method in which to classify players in today's NBA.

Both Alagappan and Shea presented a problem statement that still holds true today. The traditional five player positions incorrectly oversimplify the skill sets of NBA players. Simply pigeon-holing players into one of five positions does not accurately define a player's specific skill set. Moreover, the misclassification of a player's position may lead teams to waste resources on developing draft picks that do not fit their systems.

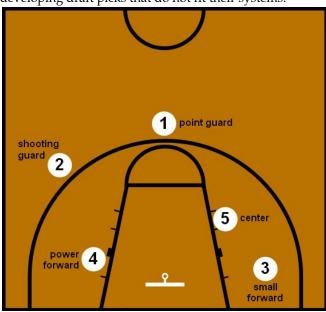


Figure 1. Traditional five position classification

Coaches and scouts in the NBA already recognize players as having skills that exceed their predefined positions and have even come up with alternative position names to describe players such as the combo guard (i.e. a player that combines the attributes of a point guard and shooting guard) or the swingman (i.e. a player that can play both the shooting guard and small forward positions). Using machine learning, my goal is to uncover the positions that are intrinsic to today's NBA players and classify players with a position that best encapsulates their skill sets.

2 LITERATURE REVIEW

In the "From 5 to 13: Redefining the Positions in Basketball", Muthu Alagappan pointed out that the traditional way of classifying the NBA players into 5 categories incorrectly oversimplify the skill sets of NBA players and the misclassification would lead to waste resources. He has found a new way of classifying by topological data analysis of data.

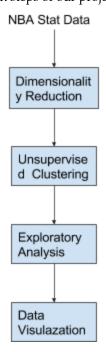
In the "Periodic Table of NBA Elements", Stephen Shea has proposed an alternative method called "Spatial Tracking" to classify the players.

Inspired by the previous work, we want to put forward a new way of classifying NBA players based on their skill sets by machine learning.

3 PROBLEM FORMULATION

We intend to classify NBA players with a position that best encapsulates their skill sets using machine learning.

Here is the main steps of our project:



1

4 DATA DESCRIPTION

We Scrape our dataset from <u>Basketball-Reference.com</u>, since there are several stat metrics in this website, we finally decide to use Per-100 Possession and Advanced Metrics to best define a player's ability. Here, we don't use cumulative statistics since those kind of metrics tend to inflate players with much longer career length. The two metrics are shown in figure 2.



Figure 2. Two stat metrics used in this project

The whole dataset we used contains 486 nba players and 45 features (i.e dimensions) in 2016-2017 season. Even though the size of dataset is really small, our goal is to classify various types of current players.

5 Proposed Solution

5.1 Dimensionality Reduction

In our project, each dimension is represented by a player's statistics(i.e. 3P%, FG%, PER). To get a sound result, we decided to reduce the total dimensions for each data point by obtaining a set of principal component. Here, since we want to implement data visualization later, we choose to get the first two principal component to build a two-dimensional project space.

For implementation part, we compared two dimensionality reduction method - PCA and LDA(Linear Discriminant Analysis) to find a better approach. We found that PCA captured 49.43% of the data and LDA captured 65.27% of the data. As a result, LDA is a better way to achieve dimensionality reduction.

5.2 Cluster the dataset

The next step for our project is to cluster our projected data by using unsupervised classification method-K-Means. So that we need to find a best parameter K which can get the best performance when we calculating the distance in the same cluster and the distance between each different clusters.

In our project, we decided to use silhouette score for evaluating different value of K(we tried different K from 5 to 15). Silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a high value indicates that an object is well matched to its own cluster and poorly matched to neighboring clusters. If the clusters have a high value, then the clustering configuration is appropriate. We can see the performance of different value of K from figure 3.



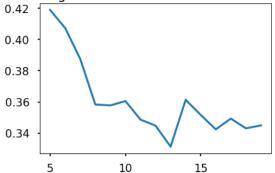


Figure 3. Silhouette Score for different value of K

Finally we decided to set K = 8, and the clustering result shown in figure 4.

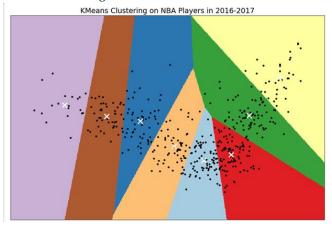


Figure 4. Clustering result

5.3 Important feature extraction

Principal Component Analysis (PCA) is a common feature extraction method in machine learning. The algorithm finds the eigenvectors of a covariance matrix with the highest eigenvalues and then uses those values to project the data into a new subspace of equal or less dimensions. In feature extraction, PCA reduces the number of features by constructing a smaller number of variables that capture a significant portion of the data found in the original features. Using PCA, we identified the most important features in order to define each cluster.

6 RESULT

6.1 Types of NBA players

After getting the important features for each cluster, we can define each cluster in some common words. Here are the names we defined for each cluster:

- 1. Versatile Center
- 2.Scoring Wings
- 3. Versatile Forwards
- 4.Offensive Centers
- 5.Floor Generals
- 6.Shooting Wings
- 7.Defensive Centers

8.3-and-D Wings

6.2 Exploratory Analysis

In this section, we intend to use two clusters (i.e Floor Generals and Defensive Centers) as examples.

1. Floor Generals

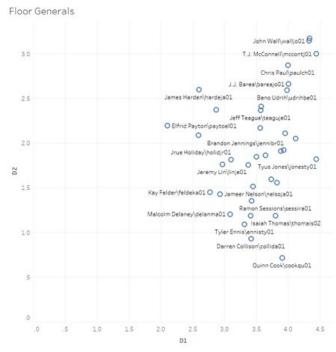


Figure 5. Data Visualization for cluster Floor General

The notable players for this cluster contains James Harden, Chris Paul and John Wall. We next take a look at 10 most important features for cluster of Floor General:

	Feature	Importance	Cluster Average	League Average
0	BPM	0.223549	-0.9 <mark>91429</mark>	-1.155631
1	PER	0.221773	15.282857	13.533559
2	VORP	0.219615	0.954286	0.677252
3	WS	0.214771	3.545714	2.822297
4	OBPM	0.211280	0.348571	-0.995495
5	WS/48	0.208305	0.087771	0.085556
6	ows	0.204775	2.188571	1.464865
7	PTS	0.191846	21.617143	19.915090
8	FT	0.189831	3.865714	3.227703
9	FTA	0.187174	4.760000	4.298198

Figure 6. Top 10 important features for Floor General

Here we can find that the top 3 important stats for this cluster are BPM(Box Plus/Minus Value), PER(Player Efficiency Rating) and VORP(Value Over Replacement Player). These three features define the importance of a player to his team, which indicates those kind of players could really control the whole team as well as the game.



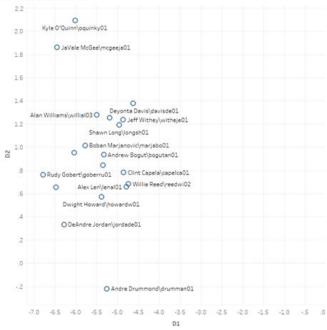


Figure 7. Data Visualization for cluster Defensive Center

The notable players for this cluster are Dwight Howard, DeAndre Jordan and Andre Drummond. The natural position of these players are center and they make more contribution on the defend side.

TOV%	0.172287	44.457005	
		14.157895	12.789414
PF	0.162561	6.542105	4.474775
DRtg	0.122652	103.052632	108.979730
DBPM	0.098983	2.136842	-0.162162
AST	0.071157	2.036842	4.261261
AST%	0.045135	6.368421	13.215991
3P%	0.025636	0.084474	0.287741
BLK%	0.023664	4.678947	1.666892
TOV	0.018149	2.615789	2.666667
STL%	0.002427	1.373684	1.563063
	DBPM AST AST% 3P% BLK% TOV	DRtg 0.122652 DBPM 0.098983 AST 0.071157 AST% 0.045135 3P% 0.025636 BLK% 0.023664 TOV 0.018149	DRtg 0.122652 103.052632 DBPM 0.098983 2.136842 AST 0.071157 2.036842 AST% 0.045135 6.368421 3P% 0.025636 0.084474 BLK% 0.023664 4.678947 TOV 0.018149 2.615789

Figure 8. Top 10 important features for Defensive Center

Here we can find a interesting point for the important features of Defensive Center, the most significant feature is the TOV%(Turnover Percentage). For me view, it is most likely because those kind of players are not good at ball handling so that they may have higher possibility to turn it over.

7 Conclusion

In this project, we classified NBA players by their feature statistics into natural clusters which better match their skill sets than traditional five positions classification. The clusters that the algorithms have generated also indicates which features are most important to a player and group them in such a way that is easily interpretable. By using this classification method, the NBA team and coach may find it easier to get to know the different play style of their players.

For future work, we plan to use more statistic metrics in order to find the combination of those metric which can define the player's skill set most precisely. Also, we can add more history players to check how the playing style of entire NBA league is changed during these years. In addition, a more in-depth exploration into each cluster may reveal insight into how teams can scout and develop the type of player they want most.