Defining New NBA Player Positions

Neel Parekh
np423@cornell.edu

Teja Arikati
ra597@cornell.edu

Abstract

The current 5-position categorization of NBA players is outdated. In this work, we propose a method to redefine positions based on the performance statistics (which is a proxy for his skills, attributes, and play-style) of a player rather than his physical characteristics. We begin by exploring the group of players traditionally labeled as Small Forwards, Power Forwards, and Centers (i.e. "Big Men") as the change in their skills has been most dramatic. We begin Phase 1 of this analysis by performing a K-Means clustering of players based on their performance statistics. We observe that three clusters best describes the data by looking at the percentage of variance described by the cluster compared to the number of clusters. Further steps are also listed.

1 Introduction

1.1 Background

The National Basketball Association (NBA) gameplay is consistently evolving. As certain player styles of play fall into or out of favor, kids who play the sport start picking up new skills modeled after their favorite players and often refine these skills to match their specific attributes. These new skills eventually make it to the NBA and leave coaches with the option to change their offensive or defensive thesis with evolving star player skills (e.g. Stephen Curry has inspired a generation of fast-passed, passing-heavy basketball focused on the 3-pt shot). Team play styles are redefined and co-evolve with individual player roles.

At the advent of basketball, people chose to label players into 5 categories (chosen primarily to match the fact that only 5 players are allowed on the court from a single team at any given time). This structure has lasted the test of time although the original roles assigned to each of the 5 positions have not. For example, players labeled Power Forwards (PFs) used to operate on offense mainly within 10 ft of the basket, exemplified by greats like Charles Barkley, Karl Malone, and Kevin McHale in the 80s/90s. In the early 2000s, we observed a surge in PFs who could shoot from outside of 15 ft (e.g. Tim Duncan, Dirk Nowitzki, Kevin Garnett). Finally today, there are many PFs who can shoot the 3-point shot (e.g. Kristaps Porzingis, Nikola Mirotic, and Kelly Olynyk). The PF label is often assigned based on the physical attributes (i.e. height, weight, speed, etc) of the player. The PFs of today neither physically nor through game style represent the same group of PFs that existed in the 1990s.

1.2 Objective

It is now time we redefine how we classify player roles. We can no longer simply lump all players together into 5 categories and expect an accurate representation of the skills in the category. The player skills and usage in each category are too varied within categories and too similar between categories. Instead, we propose replacing the 5-position system with a multi-position system based entirely on player usage and performance on the court.

By watching the game, we theorize that the ultimate number of positional categories will fall in the 8-15 range. This should cover the different combinations of skills that we observe, however the problem is ill-defined. Rather than choose a set of target categories, our work takes the approach of an unsupervised problem. Categories are chosen based on spread of the data rather than preconceived notion of likely spread of the data.

1.3 Dataset

The data we use is per game player statistics derived from stats.nba.com. The data is scraped from the website by taking the JSON response from the NBA Stats servers for each player ID, which we in turn scraped from the NBA Stats website as well. This code is provided in the references. We also noticed that the NBA servers do not return traditional positions (perhaps a validation of our hypothesis that positional categorization is meaningless). In order to get player traditional positions, we use data from http://www.espn.com/nba/salaries that includes positional data, and we merge this with the NBA Stats data on player names to create the final data structure for this phase of the project.

As player roles can often change from season to season if they switch teams (e.g. Kevin Durant) or develop new skills (e.g. Victor Oladipo), we chose to only keep the data from the most recent season (the just-completed 2017-2018 NBA season ended last week) in order to reduce the affect of positional changes between seasons. We also include only the simple "box-score metrics" in the first phase of our analysis. We normalize the statistics both by

1. minutes played, in order to account for aggregate totals discrepancies between similar players who played different numbers of minutes per game.

2. games played, in order to account for aggregate totals discrepancies as a result of games missed due to injury.

2 Methods

2.1 Baseline Model Description and Experimental Setup

We began by performing tSNE to project data into two dimensions prior to clustering. This is true for every plot included in this discussion. We then visually observed the game styles of players that are traditionally classified as Centers (C), Power Forwards (PF), and Small Forwards (SF). In order to understand how skills and attributes are spread through this basket of players we explored the features (i.e. per-minute normalized performance statistics) available in our dataset for each player in this group. We removed features that we knew (via basketball knowledge) to be entirely irrelevant to our discussion.

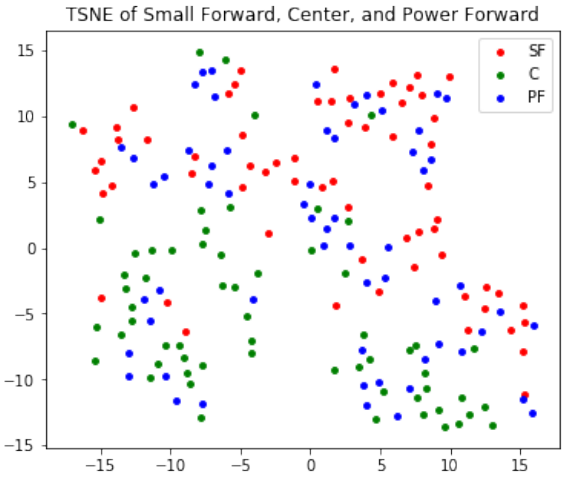


Figure 1: Original 3 Positions in tSNE

The mission is to understand how the data is clustered, so we planned to use k-means clustering to group these players into k categories. In order to understand how many clusters to keep over a range of small $k|k \in [2, 5]$, we observe the percentage of variance explained within a cluster.

3 Results and Analysis

3.1 Results

Visually, there can be cases made for any of the clustering from k = 3 to k = 5. Looking at the elbow chart comparing variance described by clusters to the number of clusters (i.e. a common variant of the silhouette coefficient), there is an inflection point at number of clusters = 3. The increased capturing of variance by the clusters starts to increase less rapidly at this point. K=3 for the number of clusters is chosen for the baseline.
Looking at the tSNE graph with the 3 known roles defined, there seems to be 2 clusters which Centers can be defined as. The PF and SF positions both seem to have no specific structure except that the SF position has less overlap with the Center position.

3.2 Analysis

When comparing the K-means clustering to the actual classes (SF,PF,C), stark differences can be found. This provides evidence that "traditional" roles have significant overlap with each other in today's NBA. The PF and C positions both seem to have significant overlap when looking at the data and this aligns with the current trends of the SF and C positions having various shared roles.

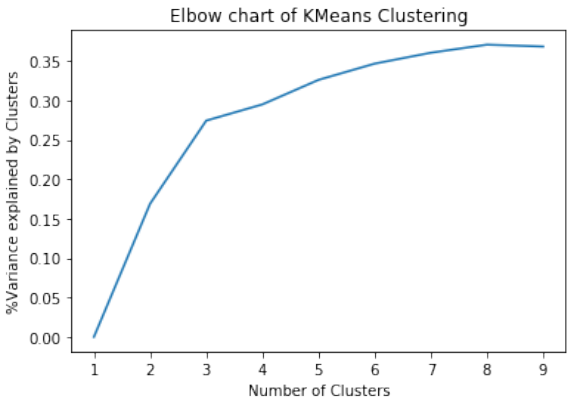


Figure 2: Inter- and intra-class variance for each cluster

For next-steps, correlations of variables and further feature engineering will be conducted in order to determine what variables drive the axes in the tSNE graph. Other forms of dimensionality reduction will be looked at, different forms of clustering will be looked at, and Regression analysis will be conducted as well. Using different sorts of models with different types of features, hopefully a better sense of positioning in today's will be obtained.

3.3 Extended Results

K-means was a good baseline to run for unsupervised clustering of the data. However, other clustering algorithms may provide better insight in separating the roles. For these next steps, we decided to run a Gaussian Mixture Model and Spectral Clustering. The Gaussian Mixture Model for clustering is very similar to K-Means however this algorithm doesn't assume equal circular variance for all the classes. It also has the benefit of being a "soft" classifier meaning there is a probability of each point belonging to each cluster. This allows analysis of class assignment at the edges and analyze cases which may be a mixture of more than one cluster.

The next case looked at was Spectral Clustering. Spectral Clustering is similar to kernel methods where it tries to lift the points into a higher dimensional space and then try to find a separation of the points in this space. This is why the clusters seem to have overlap: The separation between classes in the 2D case are not linearly/polynomially separable.

Gaussian Clustering is quite similar to K-means, the value of this method comes from being able to see the probability of each point belonging to each class. Spectral Clustering looks the most different from the other two clustering. There are points that exist primarily in one class that are actually classified to another class. This is explained by the non-linear separation boundaries between classes.

4 Analysis

For the analysis, we decided to look at the different statistics for each of the different clusters. We found that Cluster 2 exhibited higher volume shooting statistics compared to Cluster 0 and Cluster 1. When looking between Cluster 0 and 1, there doesn't seem to me much difference except that Cluster 0 has better 3 point shooting and a better plus minus score with Cluster 1 having a negative plus minus score. Both Cluster 0 and 1 have higher rebounding scores and fouls than Cluster 2.

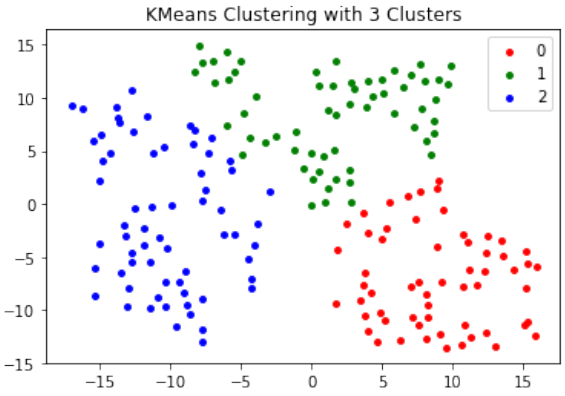
What does all this mean? We can boil this down to offensive and defensive characteristics. The offensive characteristics are more obvious in this case; shooting is the most direct characteristic for offensive abilities. Defensive characteristics are harder to gauge. We will be looking at rebounding and fouls to be indicators of more defensive capabilities. If we were to use these metrics, we can see that Cluster 2 players are more offensively oriented compared to the other two clusters. Cluster 0 players, on the other hand, are more defensively oriented. Cluster 1 seems to capture poor performing players with both shooting and defensive statistics being lower than the other two and cluster 1 also exhibits a negative plus-minus.

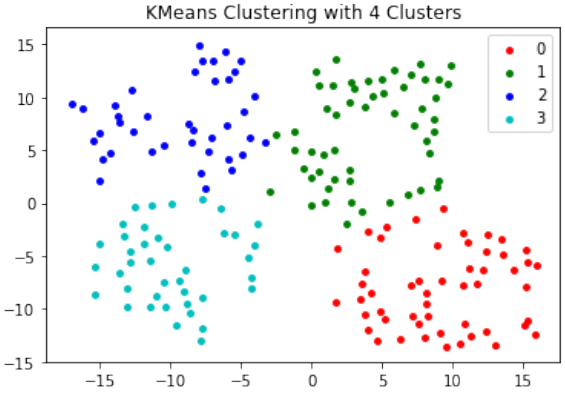
5 References

- Our Code (Jupyter Notebook) - V1
- Our Code (Jupyter Notebook) - V2
- Alagappan, Muthu

6 Image Appendix

See next page for images.





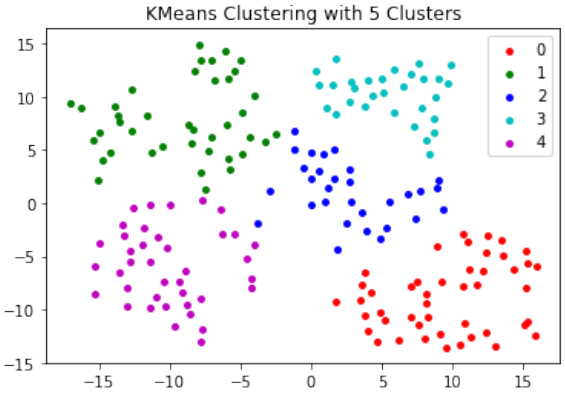
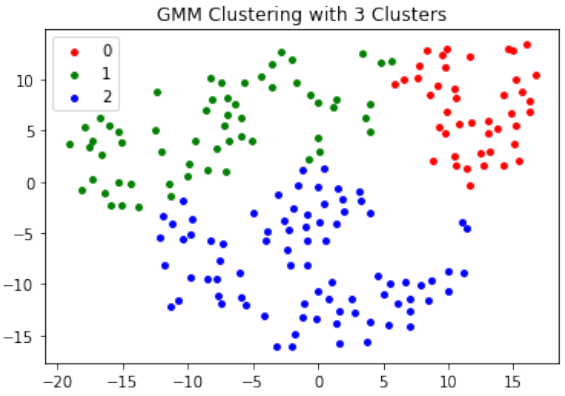


Figure 3: Clustering with K-Means



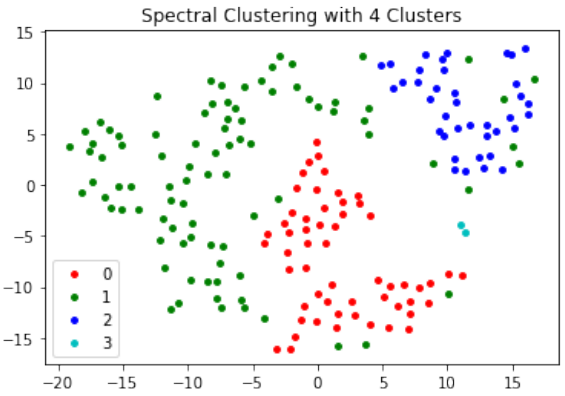


Figure 4: Clustering with Gaussian Mixture Model and Spectral Clustering