A Statistical Analysis of the Positions in Basketball: K-means Clustering on NBA players

Mitchell Hersey
University of New Hampshire
mah1021@wildcats.unh.edu

Abstract

Conventional wisdom dictates there are five positions in the NBA and basketball as a whole. These are the point guard, shooting guard, small forward, power forward, and center. However, the rise of 'positionless basketball', 'stretch bigs', and 'unicorn'-style players has dramatically altered the traditional ideas of these positions. While historically these positions denoted discrete size, skillsets, and responsibilities, this seems to no longer be the case for any avid watcher of the NBA. This paper aims to make use of the dearth of NBA statistics available and try to group players through the use of unsupervised clustering into positions based solely off of on-court production as opposed to traditional tactical thinking. Through the use of the K-means clustering algorithm, it is revealed that the shooting guard and small forward do not display any significant difference in production. The same is also revealed of the power forward and center. However, the clustering also reveals the existence of groupings of players that, while not sharing the same traditional positions, they do display similar on-court production to the point that they cannot be properly grouped with their role-player peers. Ultimately, the clustering affirms the belief that the NBA is a the most star-driven team sport as a star-players traditional role does not dictate their on-court production in any meaningful way.

Introduction

Traditionally, basketball is played with players grouped into five main positions. These are the point guard, who is generally the smallest on the court and who's main responsibility is running the offense and dictating plays in real time, they are often referred to as the 'floor general'. The shooting guard is the larger of the two guard roles and their main job is to score the basketball whether that be through 3-pointers (Kyle Korver) or slashing (DeMar DeRozan). The small forward is perhaps the least defined of all major positions but often will find themselves in a role similar to the shooting guard (Joe Ingles) or perhaps as a playmaker of their own (Magic Johnson). Like the small forward, the power forward often lacks definition aside from being the second tallest player. They can provide long range fire power (Dirk Nowitzki) or rim protection (Kevin Garnett). The center was historically the most dominant position in basketball, although not true anymore, their main responsibilities are rebounding and rim protection.

Perhaps for much of the NBA these positions are rather apt and the more team focused basketball of the 50s and 60s can be accurately divided into these five positions. However, as the NBA exploded in popularity starting in the 80s, the sport became more and more star-driven with Larry Bird and Magic Johnson leading the way for the global cultural force of Michael Jordan. The question then becomes, do these superstar players really play the same game as those players in the early days of the NBA? Michael Jordan was the best player to ever play shooting guard, but does his statistical output even resemble a shooting guard from 1952? If not, is it really true that these players play the same position?

One possible approach to this question would be through the use of a classification algorithm. The issue with this method though is its susceptibility to human bias. Classification algorithms rely upon training data that the user specifies. However, in the case of NBA positions, what even is the correct training data? A point guard from 1950 is not necessarily a better example of the position than a modern one.

Another approach would be to try and group players through regression. The issue in this case is that regression relies upon independent and dependent variables. In the case of NBA players, what even would a dependent variable be? This approach then becomes a struggle to try and justify the relationship between statistics when there really is no objective answer.

In order to address this question without the implicit bias of a human observer, we can try and group players strictly through their statistical output. The most obvious way to do so is through the clustering method of machine learning. With clustering, a player's season can be grouped solely through what the box score says they did in comparison to their peers as opposed to a decision the coach made.

Through clustering the traditional NBA positions can be affirmed through statistical similarity or new groupings can be observed based solely off of quantitative data.

K-means Clustering

A simple but effective method for determining the statistical similarity of multi-dimensional observations is through the K-means clustering algorithm. Specifically, Lloyd's Algorithm which treats every observation as a point in multi-dimensional space.

The algorithm begins with a defined number of groups (clusters) K to partition the data into and a dataset D of n observations. K random observations are selected as the initial centers of these clusters. Then, for every observation in D, determine the nearest cluster centroid by Euclidean distance and assign it to that cluster. After every observation has been assigned, the clusters are re-adjusted to be the mean of the observations assigned to them. This process repeats until the clusters no longer re-adjust and the dataset has been partitioned (Lloyd 1982).

While K-means clustering is a simple algorithm it often performs quite well. The major decisions to be made are the dimensions to use in each observation, the number of clusters, and the initial centers of each cluster.

At the most general level, the data should be partitioned into at least two separate groups, backcourt and frontcourt players. Height is incredibly important in basketball and it results in frontcourt players generally getting more than double the rebounds of their backcourt players.

Progressing further, backcourt players will likely be divided into groups of wings and ballhandlers. The point guard, due to ability and the sheer amount of time they have the ball, accumulates many more assists than other backcourt players.

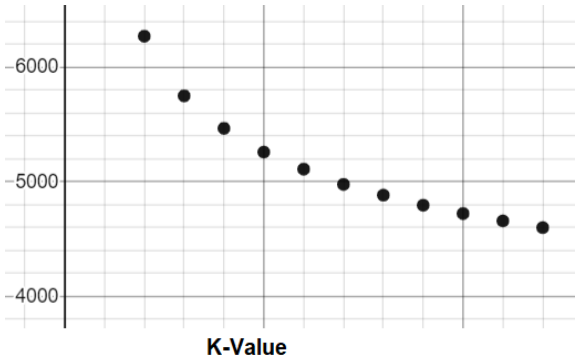
Ideally, the Euclidean distance between these group (taking other statistics into account) should portray these types of players as their own disparate clusters.

Observations and Dimensions

Before any clustering can be considered, the data to use must be decided upon. Our observations are a player's season with a team. This then treats a player that played with two teams in a season as two discrete observations.

However, for a season to be considered, it must meet some criteria. Seasons are only considered in which a player's minutes exceed 400. This provides a baseline for the player's performance that can be projected onto a 36 minute per game workload. Speaking of this, raw box stats are projected onto a per 36 basis which allows players that only play a few minutes per game to be on equal footing with a player in a starting role. This is a standard practice in the basketball statistics community. The counting stats are again normalized between 0 and 1 so that no dimension (points in particular) will be too impactful on the data.

As far as the dimensions being used, that will depend upon the sample to be studied. There are two separate clustering procedures to be done based upon the statistics available. The dimensions used in every clustering are **points**, **assists**, **free throw attempts**, free throw percentage, **three-point attempts**, three-point percentage, **field goal attempts**, field goal percentage, and true shooting percentage. The fields marked in bold are counting stats which are normalized to a per 36-minute basis. These statistics are available starting in 1950, however minutes were not recorded until 1952. Therefore, the largest possible clustering can be done with the data available for seasons from 1952 to 2017 in 9-dimensional space.



A second clustering is done with even more dimensions to try and create a better idea of a player's playstyle through their statistical output. In addition to the statistics above, beginning in 1978 onwards, a clustering can be done with **rebounds**, **turnovers**, **steals**, **blocks**, usage rate, **two-point attempts**, and two-point percentage. Usage rate is a statistic that measures the percentage of offensive possessions that a player ends the ball with. Whether that be a shot or turnover. It generally is used to measure a player's 'ball-dominance' and is a good indicator of how often the player is touching the ball. This clustering in 16-dimensional space can be done from 1978 onwards and provides more insight into what is generally considered the 'modern' NBA post ABA-merger and the advent of the three-point line.

Choosing K

Ultimately, clustering is most dependent on the K-value which dictates the number of clusters to use. K should be a middle-ground between usable clusters of observations such that broad statements can reasonably made on them, the most general being two groups. It must also minimize the error between the cluster centroid and an observation which would mean a cluster for every point of data.

There are many ways to determine the optimal value for K that can rely on both quantitative and qualitative decisions. In this case, to most directly address the central question of whether players neatly cluster into five positions, the

obvious K-value should be five. However, not all clustering experiments have the benefit of such an obvious K-value and must rely on techniques to determine it.

Elbow Method

A strictly qualitative method known as the elbow method depends on a measure of the difference between an observation and its centroid. This value should decrease to 0 as K increases to N such that when every data point has its own cluster the cumulative error is 0.

In this case the sum of squares error can be used as a sum of the total error (SE) between centroid and observation with the error simply as the Euclidean distance. What the elbow method does is to try and locate the 'elbow' or 'knee' in the graph of SE vs K-value. It tries to visually identify the aforementioned middle ground where a higher K-value provides less benefit.

The graph depicted above displays the 'elbow' where the steeper SE becomes more horizontal. Obviously, there is some elbow in this graph, but the short-coming of the elbow method is that the specific K-value is often open to interpretation. I would identify it as a K-value of six, but one could reasonably say the value is anywhere between five and 8 (Thorndike 1953).

Silhouette Model

A more quantitative approach to determining the correct K-value is using some metric for the quality of a cluster and determining which K-value produces the clusters with the best metric.

In this case, the silhouette value was chosen as a measure of cluster quality. The silhouette value attempts to determine the similarity of an observation within its own cluster compared to its nearest neighbor cluster. It does this, by determining the average Euclidean distance between an observation i and all other observations j within its own cluster, C_i. The distance is d(i,j).

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

Then, the observation is compared to members of all other clusters and the most similar cluster is determined to be the nearest neighboring cluster, C_k.

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

Ideally, a(i) should be very small, indicating a closely knit cluster and b(i) should be large, indicating even the nearest neighboring cluster is far away.

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i). \end{cases}$$

The silhouette value s(i) then will lie somewhere between 1 and -1 with 1 indicating very good clustering and -1 would indicate that the observation should actually be within the neighboring cluster. Obviously K-means would never have a value below 0, but 0 itself would indicate that the observation sits on the border of the two clusters which is decided arbitrarily according to the K-means algorithm (Rousseeuw 1986).

The silhouette model then makes use of the average silhouette value for every observation to determine an optimal K-value.

1978-2017 Clustering Average Silhouette Values	
K-Value	**Average Silhouette Value**
2	0.605953
3	0.604728
4	0.597164
5	0.586858
6	0.57814
7	0.578231
8	0.573567
9	0.567595

The silhouette values depicted above obviously do not present a leading candidate for the optimal K-value. Ultimately clustering quality can depend upon the initial observations used for seeding which may account for the slight variation in silhouette values. However, what this data does tell us is that while by no means perfect, the data set can be used decently well in a clustering algorithm. None of the values drop below 0.5 and nothing even comes close to that dreaded 0 silhouette value.

Cluster Seeding

With the K-value, data, and suitability for clustering decided upon, there is one more consideration to be made before the K-means algorithm can be run. Ultimately the most important value in determining a cluster is the initial observation chosen for it. If this observation is some significant outlier, the cluster may only consist of that single observation and provide no insight into the clustering of the dataset as a whole.

To account for this, we can run the K-means algorithm multiple times with different starting points. Then, the clustering with the best average silhouette value can be considered the best clustering of the data. Luckily, this dataset doesn't contain any significant outliers once counting stats have been normalized to per 36-minutes and silhouette values vary be less than 1-standard deviation throughout hundreds of runs of the algorithm.

Results

The core goal of this project was to determine whether or not players cluster through their on-court performance into their specified position. The answer to this is, sort of.

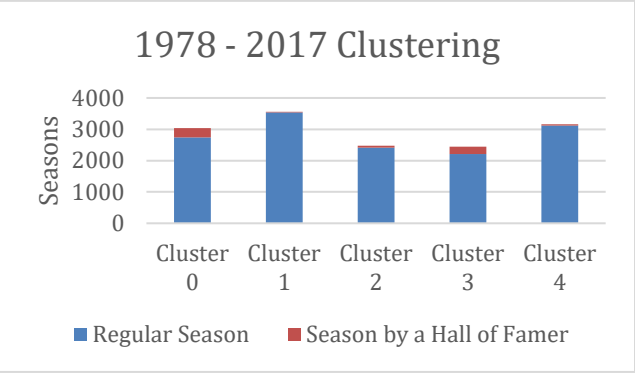
When looking at the clustering from 1952-2017 it becomes clear that over the course of the NBA, basketball has changed so much that the traditional positions don't mean nearly as much as they used to but that they also never meant all that much. Basketball can be more generally divided into three main positions. Ball handlers who run the offense, wings who score from range and play defense accumulating steals, and big men who score efficiently from near the basket and accumulate blocks and rebounds. Any avid NBA watcher could tell you that the distinction between a shooting guard and small forward is quite vague and the same can be said of power forwards and centers.

1952-2017 Clustering Position Distribution					
Cluster	**PG**	**SG**	**SF**	**PF**	**C**
1	19.6%	17.7%	17.2%	19.0%	20.3%
2	26.4%	**36.8%**	**28.0%**	12.0%	1.7%
3	9.0%	20.9%	23.3%	22.0%	20.0%
4	1.3%	3.3%	10.9%	**34.1%**	**49.9%**
5	**43.6%**	21.3%	20.6%	12.9%	8.1%

The largest clustering does appear to support this assumption as shooting guard and small forward are grouped together while power forward and center are as well. The truly interesting piece of analysis in my mind are the populations of the clusters that are not dominated by one or two positions.

A Star Driven League

Basketball is unique among team sports in the ability for one player to lead their team to victory. In many ways it would appear that a Lebron James or Michael Jordan are simply playing a different game than everyone else on the court.



It is these types of players fit into those two extra clusters. Based strictly off of on-court performance basketball is a game divided between stars and role players. Of the 666 seasons by a Hall of Famer between 1978 and 2017, 80.4% of them belong to the 'star' clusters. According to the clustering, a role player and star player in the same position are not peers but actually playing different roles entirely.

Ultimately, the clustering confirms what I believe to be the greatest part of the NBA. At its core, basketball will always be a team game, but the ability of a single player to be truly amazing is undeniable both qualitatively and statistically.

Extensions

While I am happy with my results and program, there are certainly some areas I would like to explore related to this. It would be interesting to examine how other clustering algorithms treat this same data set. Algorithms like EM, K-medoids, Fuzzy clustering, etc. take different approaches to grouping the data and may provide new observations.

Positionless Basketball

There is a lot of talk about 'positionless' basketball. The idea that players are no longer limited to the traditional responsibilities of their position and instead their own talent level. The rise of stretch bigs who can shoot threes and talented wings capable of running the offense.

Fuzzy means seems best suited to quantifying the hybridization of positions as it accounts for the weight of an observation to multiple clusters.

Shortcomings of Clustering

What I feel I learned the most from this project however was the inherent issues with clustering. My application was unique in that there is a hypothetical K-value to compare to, but most applications don't have that benefit.

Clustering is supposed to try and remove the human bias in grouping but ultimately with the emphasis on K-values, it is extremely hard to truly remove oneself from that bias. While there are a range of statistics to try and determine an optimal K-value it still seems to be a field without any true answer. That is not even to mention the inherent biases present in the selection of attributes for an observation.

I chose statistics that I believe tell a complete story of a player's on court performance, but they are still rife with my own personal beliefs. I could have just as easily chosen effective field goal percentage instead of true shooting percentage, but the choice ultimately came down to my own subjective favor of TS%. Choosing both then brings the issue that the two are almost always correlated and now my data would have a bias for scoring efficiency.

My last concern is the use of per 36-minute stats. While they are generally the best way to try and plot a player's performance as a starter, nobody would claim they are perfect. Ultimately there is a reason a coach is not playing a player who performs extremely well in limited time. A notable

example is Boban Marjanović who posts all-star level stats per 36 but cannot physically sustain that pace of play.

Clustering is incredible in theory, but this project has highlighted how hard it is to escape my own inherent biases and handle the data in an objective way.

References

S. Lloyd, "Least squares quantization in PCM," in IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, March 1982, doi: 10.1109/TIT.1982.1056489.

Thorndike, R.L. Who belongs in the family?. Psychometrika 18, 267–276 (1953). https://doi.org/10.1007/BF02289263

Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.