Understanding Player Positions in the NBA

Luke Lefebure
Stanford University
lefebure@stanford.edu

December 12, 2014

Abstract

To better understand the meaning of positions in the NBA, I train several models to predict a player's position from his stats, such as average speed on the court and number of three point attempts per game. The NBA and ESPN use different position labels, so I compare the performance of my models in predicting these labels separately. Ultimately, I hope that this analysis can determine how well position labels describe NBA player roles and identify the defining characteristics of each position.

1 Introduction

Basketball is a unique sport in that there are no position specific rules or statistics. All players are traditionally measured by the number of rebounds, blocks, and steals they get on defense and the number of points they score, passes they make, and assists they get on offense. Therefore, position labels can be somewhat ambiguous: the NBA classifies players as either a Guard, Guard-Forward, Forward, Forward-Center, or Center, and ESPN classifies players as either a Point Guard, Shooting Guard, Small Forward, Power Forward, or Center. This has led some to do cluster analysis on NBA players to discover more and "new" positions other than the traditional five [1].

The goal of this analysis is to determine how well existing position labels actually describe a player's game. This is done by building models to predict position label from player stats. I consider two different position labeling schemes, namely those given by the NBA and ESPN, and compare model performance in predicting these labels separately. In addition, by analyzing these models, I can identify the features that are most predictive of position.

2 Data

This analysis looks at 340 players from the 2013/2014 NBA season. I use 19 features for each player to predict two separate position labels.

2.1 Response

I scraped players' NBA and ESPN position labels from those organizations' respective websites [2] [3]. Table 1 below shows the relationship between these two position labeling schemes for players from the 2013/2014 season. The middle row of this table, for example, says that among players classified by the NBA as Forwards, ESPN classifies about half of them as Small Forwards and the other half as Power Forwards.

Table 1: Relationship between NBA and ESPN Position Labeling

NBA/ESPN	Point Guard	Shooting Guard	Small Forward	Power Forward	Center
Guard	74	54	0	0	0
Guard-Forward	0	16	9	0	0
Forward	0	0	56	59	0
Forward-Center	0	0	0	19	15
Center	0	0	0	0	38

2.2 Predictors

Prior to the 2013/2014 season, the NBA installed SportVU cameras into every team's arena. These cameras capture what is known as Player Tracking data, which consists of advanced player stats such as distance traveled and secondary assists. These are the stats that I use as predictors.

Because of the novelty of these cameras, I am limited to one full season of this data. I further condensed my dataset by including only players who played at least 30 games during the season and averaged at least 10 minutes per game in order to ensure that each player had sufficient data.

The Player Tracking data contains around 100 features for each player, but much of it is repetitive. For example, points per game is included but so are total points and number of games played. Additionally, this analysis attempts to focus on style of play as opposed to quality of play, meaning that I am interested in statistics such as number of three point attempts rather than three point percentage. Table 2 below shows the features that this analysis considers, all of which are continuous.

Table 2: Features Considered

Stat Abbreviation	**Description** [2]
GP	Games Played
MIN	Minutes per game
defense.BLK	Blocks per game
defense.STL	Steals per game
passing.PASS	Passes made or received per game
passing.AST_POT	Assist opportunities per game (passes by a player to a teammate in which the teammate attempts a shot, and if made, would be an assist)
rebounding.OREB_CHANCE	The number of times per game a player was within the vicinity (3.5ft) of an offensive rebound
rebounding.DREB_CHANCE	The number of times per game a player was within the vicinity (3.5ft) of a defensive rebound
catchShoot.FG3A	Catch and Shoot three point attempts per game
catchShoot.FG2A	Catch and Shoot two point attempts per game
pullUpShoot.FG3A	Pull Up three point attempts per game
pullUpShoot.FG2A	Pull Up two point attempts per game
shooting.FGA_DRIVE	Field goal attempts per game on drives
shooting.FGA_CLOSE	Field goal attempts per game starting from touches within 12 ft of the basket, excluding drives
drives.DVS	Number of drives per game
touches.TCH	Number of times the player touches and possesses the ball per game
touches.TOP	Total amount of time the player possesses the ball per game
speed.DIST_48	Distance (in miles) traveled per 48 minutes
speed.AV_SPD	The average speed (in mph) of all movements (sprinting, jogging, standing, walking, backwards and forwards) by a player while on the court
Term	**Explanation** [2]
Drive	Any touch that starts at least 20 feet from the hoop and is dribbled within 10 feet of the hoop, excluding fast breaks
Catch and Shoot	Any jump shot outside of 10 feet where a player possessed the ball for 2 seconds or less and took no dribbles
Pull Up	Any jump shot outside 10 feet where a player took 1 or more dribbles before shooting

2.3 Standardizing the Data

Because most of the stats are given on a per game basis and not all players play the same amount of time during a game, I standardized these statistics to be reported on a per 48 minute basis. Specifically, if statistic s_p is reported as a per game statistic (e.g. blocks per game) for player p and player p plays on average m_p minutes per game, I transform the statistic as $\frac{s_p}{m_p} * 48$.

To allow for more interpretable results, I then standardized each variable as $\frac{s_p - \bar{s}}{\sigma_s}$, where s_p is the player's stat, and $\bar{s}$ and σ_s are the average and standard deviation of that stat respectively across all players.

3 Models

I explored three different models that perform multi-class classification: Multinomial Logistic Regression, Linear Discriminant Analysis, and Support Vector Machine.

3.1 Multinomial Logistic Regression

For a classification problem with k classes, Multinomial Logistic Regression estimates $k-1$ coefficient vectors, $\theta_1, ..., \theta_{k-1}$ and defines θ_k to be the zero vector. These coefficient vectors determine the probability that a given sample x belongs to any of the k classes. Predictions are assigned to the class with the highest probability. Formally, if $y = 1, ..., k$ are the k classes:

$$P(y = j|x; \theta) = \frac{exp(\theta_j^T x)}{\sum_{i=1}^{k} exp(\theta_i^T x)} \; j = 1...k-1$$

$$P(y = k|x; \theta) = 1 - \sum_{i=1}^{k-1} P(y = i|x; \theta)$$

In this analysis, coefficients are estimated using the multinom function in R's nnet library. The utility of these coefficients in understanding how this model makes predictions is explained later in this paper.

3.2 Linear Discriminant Analysis

LDA models $p(x|y)$ as a multivariate normal distribution, and a strong assumption is that all classes share the same covariance matrix, Σ. Formally:

$$x|y = i \sim N(\mu_i, \Sigma)$$

This could be problematic in this dataset because some features may vary a lot more for certain positions than others. Centers, for example, rarely drive the ball because they spend most of their time on offense under the basket, so the drives.DVS feature should have a relatively small variance for Centers. However, there should be a lot more variance for this feature among Guards because some Guards drive the ball while others are primarily perimeter shooters. In future analysis, I will look at Quadratic Discriminant Analysis, a generalization of LDA where each class has a separate covariance matrix, in order to hopefully remedy this problem. I use the lda function in R's MASS library to implement LDA.

3.3 Support Vector Machine

To implement SVM, I use the svm function in R's e1071 library. This is R's implementation of LIBSVM. For multi-class classification with SVM, LIBSVM uses the "one-against-one" approach in which $k(k-1)/2$ classifiers are built, each of which trains on data from only two classes [4]. I experimented with several different kernels, including linear and radial kernels, but I only report the error from the linear kernel because this performed best.

4 Results

I assessed the performance of these models in two ways. First, I trained on a random 70% split of the data (226 players), tested on the remaining 30% (114 players), and reported the training and test error. Second, I performed 10-fold cross validation and reported this estimate of test error. The error metric used is the misclassification rate. Tables 3a and 3b summarize the results.

Table 3a: Results on NBA Position Labeling

Model	Training Error (70%)	Test Error (30%)	10-Fold CV Test Error
MLR	.199	.388	.344
LDA	.207	.379	.356
SVM (linear kernel)	.199	.282	.347

Table 3b: Results on ESPN Position Labeling

Model	Training Error (70%)	Test Error (30%)	10-Fold CV Test Error
MLR	.216	.379	.376
LDA	.212	.318	.329
SVM Linear Kernel	.212	.309	.335

4.1 Error Analysis

A confusion matrix gives a better idea of the performance of these models because certain misclassifications are better/worse than others. Basketball positions fall roughly on a spectrum with Guards on one end and Centers on the other. Therefore, we absolutely do not want to see a Point Guard classified as a Center, but it is understandable to have a Point Guard classified as a Shooting Guard. Table 4 shows a confusion matrix of the ESPN position predictions of Multinomial Logistic Regression on a 30% test set. The first column, for example, says that among Point Guards, this model classified 24 as Point Guards and 2 as Shooting Guards.

Table 4: Confusion Matrix for MLR Test Predictions

Predicted/Actual	Point Guard	Shooting Guard	Small Forward	Power Forward	Center
Point Guard	24	2	0	0	0
Shooting Guard	2	15	8	1	0
Small Forward	0	4	11	5	0
Power Forward	0	1	3	16	10
Center	0	0	0	5	7

5 Discussion

SVM outperformed the other models when training and testing on a 50/50 split (results not shown), but the other models caught up quickly as demonstrated by the errors for the 70/30 split in Tables 3a and 3b. All of the models ultimately produced remarkably similar levels of accuracy on both position labeling schemes, and none of them were able to predict position with greater than roughly 65% accuracy. This result suggests that there is a lot of overlap among positions.

This analysis can also be used to better understand the existing position labels. In Table 4, we see that there are several Power Forwards classified as Centers, two of whom are Pau Gasol and Tiago Splitter. What separates these players from other Power Forwards? By analyzing the Multinomial Logistic Regression model coefficients, we can answer this question. Both Gasol and Splitter are distinguished by their above average rebounding ability, and as a result, they both have large positive values for rebounding.OREB_CHANCE and rebounding.DREB_CHANCE. Conversely, the Power Forward coefficients for these two features are very negative in the model that uses Center as the referent group (the group for which the coefficients are defined to be zero). Referring to the equations for Multinomial Logistic Regression detailed previously, this means that positive rebounding values (above average), boost a player's probability of being a Center relative to being a Power Forward, and vice versa. This partly explains why both Gasol and Splitter were classified as Centers.

Similar analysis of these coefficients can be done for other positions as well. Point Guards have a much higher positive coefficient than all other positions for passing.PASS, meaning that if a player makes a lot of passes, his probability of being a Point Guard relative to being any other position is boosted. Point Guards, Shooting Guards, and Small Forwards all have a similar positive coefficient for pullUpShoot.FG3A while this coefficient for Power Forwards is near zero. Thus, if a player takes a lot of pull-up threes, his probability of being either a Point Guard, Shooting Guard, or Small Forward goes up relative to his probability of being a Power Forward or a Center.

6 Conclusions

I expected that I would not be able to predict position with very high accuracy. In theory, basketball positions are well-defined — Point Guards possess and pass the ball while Shooting Guards take shots — but in reality, the lines between positions are blurred. The fact that I was able to achieve similar accuracy with two different position labels (albeit with substantial overlap) reinforces this notion. These results suggest that doing a cluster analysis to "redefine" positions in basketball is justified, such as that done by Lutz [1].

However, this analysis does provide a lot of insight into what features are characteristic of existing positions. By interpreting the coefficients of the Multinomial Logistic Regression model for example, we were able to see which features distinguished different positions. This paper details only a small portion of the information about positions that can be inferred from this analysis.

7 Future

Because the Player Tracking data exists for only one full season, I was limited by a small sample size. As more data becomes available, I will be able to extend my analysis further. I also need to address problems of collinearity in the dataset. Some features, such as number of possessions and time of possession, show some correlation, and this causes these features to have unstable coefficient estimates in Multinomial Logistic Regression. Similarly, I would like to use a more elaborate feature selection strategy. Finally, I would like to try Quadratic Discriminant Analysis because this model does not have the potentially problematic assumption of LDA described previously.

References

[1] D. Lutz, "A Cluster Analysis of NBA Players," in *MIT Sloan Sports Analytics Conf.*, Boston, MA, 2012

[2] NBA. Web. 7 Nov. 2014. http://stats.nba.com/

[3] ESPN NBA. Web. 25 Nov. 2014. http://espn.go.com/nba/players/

[4] C. Chang and C. Lin. "LIBSVM: A Library for Support Vector Machines," National Taiwan University, Taipei, Taiwan, 2013.