CS91r Final Submission

Matt Goldberg

January 2, 2017

1 Introduction

Much of the work being done in the world of NBA analytics attempts to answer the problem of how an individual player should be valued. In particular, a common insight is that basketball is inherently a team sport, and the players with whom and against whom one plays can significantly alter one's traditional box score statistics. Therefore, a major problem in player evaluation is teasing apart an individual player's contribution to his team without falsely attributing contributions to poor players who merely play with great players (or against even worse players), and without overlooking great players on mediocre teams. The major breakthrough in player evaluation that addressed these issues was adjusted plus/minus[5], which controls for a player's teammates and opponents to determine what effect a player has on point differential, independent of who else is on the court. Since this framework was introduced, many others have expanded on it and adopted the mindset of breaking down plus/minus [1][8][4].

However, NBA general managers and coaches do not make decisions about a player based on his value in a vacuum; rather, they make decisions based on a player's value in the context of the team they already have. Notably, chemistry is often critical to success in the NBA, so there is a strong incentive to ensure that players fit well together[7]. Therefore, my goal is to build a principled model that, at its core, predicts the point differential between a given two lineups, based on the talent and play style of the players involved. Using this model, a coach would be able to determine how player's specific skillset would make an impact given their current roster. Moreover, a coach could determine which players on a roster would complement each other well and should play a significant portion of their minutes together.

2 Background

One related piece of literature is a paper presented at Sloan Sports Analytics Conference in 2012 by Allan Maymin, Philip Maymin, and Eugene Shen[3]. In this paper, the authors used a skills plus/minus framework to evaluate a player's offensive and defensive capabilities in scoring, rebounding, and ball-handling. Then, they used these skillset metrics to simulate games, which measured the success of the individual players involved as well as the overall lineup. While the goal of this paper is very similar to mine, they took a different approach than the one that I plan on taking; most notably, I am using clustering rather than regression to approximate play styles and I am using regression rather than simulation to evaluate players and lineups. Nonetheless, this paper is very inspiring in terms of the questions it sets out to answer, so it will surely aid me throughout my thesis. I will now lay out the basic setup of my model.

First, adjusted plus/minus is essentially a regression on play-by-play data where each possession is an observation, and the predictors are indicators for each player on the court on the home team and each player on the court for the away team. Moreover, regularized adjusted plus/minus expands on adjusted plus/minus by utilizing ridge regression instead of ordinary least squares, thus preventing overfitting for players with low sample sizes and adding stability to the problem. I will be using some form of APM or RAPM as a proxy for talent, or more precisely, for how much a player contributes to his team, regardless of his play style and independent of his teammates on the court.

Second, I will use various clustering techniques to ascertain each player's play style. More specifically, I will select features that are indicative of a player's skillset and how they play the game. This is a very common technique in basketball analytics, and it is frequently used to cluster players into "positions" outside the traditional five positions[2]. Moreover, I will apply a clustering algorithm such that each player-season is assigned to a cluster, allowing a player's play-style to change each season. This will give a proxy for what kind of player someone is, without speaking to how talented the individual is.

Finally, I will combine these two measures of each player into an APM-like regression in which the dependent variable is point differential and the predictors are proxies for each player's talent (i.e. APM) and play style (i.e., output from the clustering algorithm). More specifically, I plan to predict point differential in a game in year Y using play style and APM data from year $Y - 1$. Using this data for the players on the court, I will then predict point differential in order to ascertain the impact of different play styles and different levels of skill on point differential.

3 Clustering Approach

This section will discuss the approach to clustering data in order to deduce each player-season's play style.

3.1 Feature Selection

In selecting player features on which to cluster, it is important to select features that are as informative about a player's play style as possible. For example, while points per game might be an important statistic elsewhere, it is not particularly telling about a player's play style because it is highly dependent on how many minutes he gets from his coach. Likewise, true shooting percentage is not as informative about a player's play style than his distribution of shot distances, since the latter says more about the player's shot selection. Moreover, true shooting percentage and statistics like it are likely to be correlated with APM, which is already included in the final regression.

Keeping this in mind, I have currently selected 27 features to represent a player's play style (at least, before possible dimensionality reduction). These features are summarized in Table 1.

Area of Play	Features
Vitals	Height, Weight
FGA Distribution	Avg Shot Distance, Pct of FGA from 0-3 feet, 3-10, 10-16, 16+ foot 2-pointers, and 3-pointers
FGM Distribution	FG% from 0-3 feet, 3-10, 10-16, 16+ foot 2-pointers, and 3-pointers
Scoring Type	USG%, free throw rate, percent of 2PM assisted, percent of 3PM assisted
Rebounding	ORB% DRB%
Defense	Defensive rating, BLK%, STL%, PF/poss
Ball-Handling	AST%, TO%, bad passes per minute, lost balls per minute

Table 1: Features Selected for Clustering

There is a concern that the number of variables selected in each area of play may implicitly give different weights to each area of play, depending on the clustering method. For example, it may be the case that rebounding is overlooked due to a relatively smaller number of features representing rebounding. This is something I will keep my eye on, and it is something that dimensionality reduction may be able to alleviate.

3.2 Clustering Algorithms

So far, I tried two different clustering algorithms: first, I tried standard K-means. K-means is a clustering algorithm that iteratively assigns points to the nearest cluster center and reassigns cluster centers to the centroid of the points assigned to that cluster; the process ends when no points are assigned to different clusters. Therefore, K-means maintains the property that any point is part of the cluster with the closest center. Because K-means requires computing distances between points, it is important or each dimension to have the same units; therefore, I normalized the data before clustering by subtracting the mean and dividing by the standard deviation of each feature. This is also useful because it makes it easier to tell which players are above or below average (and to what extent), which is especially useful for less well-known features.

Second, I tried a Gaussian mixture model, a generative, probabilistic model. A GMM is similar to K-means, except it models each component as a multivariate Gaussian, meaning that different components can have different covariance matrices, and thus different shapes. Moreover, it uses soft assignments in that it models each point as a mixture of all K components; in other words, it allows a point to belong 20% to one cluster and 80% to another cluster. This may be useful because it shows to what degree a player is between two clusters, or a "hybrid" player.

I also tried to reduce the dimensionality of the feature space by applying PCA before applying clustering methods; I took the first 10 principal components, which explained about 80% of the variance in the 27 features.

Finally, I evaluated how well a clustering fit the data by attempting to maximizing the average silhouette score of the data, given a clustering[6].

4 Results

For most of the results, please see the notebook that will be submitted with this writeup, as it has both the analysis/clustering code and the results. Empirically, based on the features I selected, there seems to be around 7 clusters in the data; it seems that fewer clusters actually fit the data better, but in order to represent play styles effectively it is helpful to have smaller, more granular clusters. This selection of the number of clusters will be something that I figure out as I do the regression analysis, as the best choice for K will be the choice that allows the final regression to perform best.

5 Future Work

The next thing to work on for me is to finish up clustering experiments and then to move on to formulating APM and creating a regression. Most importantly in this is scraping play-by-play data from basketball-reference.com effectively so that I can keep track of the lineup in the game at any point in time.

Other things I need to think about as I move forward with my thesis are the questions I want to answer once I have results. For example, possible questions I may wish to answer include:

- What is each team's best lineup? This is especially helpful when starters are resting; what is the best lineup I can play without the players I need to rest? This could lead to a coaching decision model that incorporates fatigue as well as chemistry.
- Are certain types of players more valuable, or perhaps more replaceable, than others?
- Which coaches are the best at assembling rosters and playing players that complement each other well?
- Given a roster and a possible free agent/trade acquisition, how valuable is the deal in question? How well does it fit?
- Are there any notable mutually-beneficial trades that could happen?
- How have the effects of complementary skillsets changed over time?

Clearly, there are a lot of different ways I could take this project upon finishing the regression analysis.

References

[1] Jeremias Englemann. Youtube, Nov 2015.

[2] David Lutz. A cluster analysis of nba players, Feb 2012.

[3] Allan Maymin, Philip Maymin, and Eugene Shen. Nba chemistry: Positive and negative synergies in basketball. *Sloan Sports Analytics Conference*, Mar 2012.

[4] Daniel Myers. A review of adjusted plus/minus and stabilization, May 2011.

[5] Dan T. Rosenbaum. Picking the difference makers for the all-nba teams, Apr 2004.

[6] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:5365, 1987.

[7] Michael Schrage. Team chemistry is the new holy grail of performance analytics, Mar 2014.

[8] Joseph Sill. Improved nba adjusted /- using regularization and out-of-sample testing. *Sloan Sports Analytics Conference*, Mar 2010.