Identifying Strategic Patterns in NBA Positional Tracking Data
CS 229 Project Final Writeup

John Sears and Gabriel Poon

Abstract

This paper proposes a novel unsupervised learning approach to drawing insight into basketball play-calling decisions using optical data. We project inter-temporal movement of players through space onto a grid of coordinates, and use k-means clustering to identify both the common plays that teams run on a regular basis, as well as detect outlying, strategically significant plays run by teams in high-leverage situations. Finally, we use principal component analysis to validate our feature selection.

Introduction

Professional basketball is a multi-billion dollar business, and teams search vigorously for exploitable advantages that will increase their odds of winning a championship. One method teams employ is sending team employees, called advance scouts, to other cities to watch games involving upcoming opponents. The scouts diagram each play that the observed teams run, then report these back so that their own coaching staff and team can plan for how to strategically counter them.

In this paper, we will define some basketball vocabulary as follows. A *possession* is the time between when a team gains control of the ball and the time they lose it (via a missed shot, a turnover, or a foul). A *play* is the set of movements that a team employs on a possession. There are thirty teams in the National Basketball Association (NBA), and so by employing about five scouts a team can diagram about a third of their opponents games. Finally, teams run variants on a small number of plays (about 15 according to an employee of one team with whom we talked) almost all of the time, but in high-stakes circumstances (when the outcome of the game depends on one or two possessions) will switch to more scripted, novel plays that they believe the defense will not have prepared for.

Related Work

Perse et al. use optical data from basketball games to classify plays. They manually define criteria for recognizable basketball actions that are components of a play, then use the raw optical data to recognize occurrences of each type. They build semantic descriptions of plays composed of all observed actions, then train their algorithm to match these to manually defined play types. By contrast, we do not assume specific structure to actions or plays, nor do we do we require manual labeling of the data.

Haase and Brefeld use optical data for soccer games to track individual movement, then represent it as translation-, rotation-, and scale-invariant in order to find similar patterns of movement on the field, without attempting to tie them to any strategic intent. Our approach does not restrict focus to individual players, but attempts to derive meaning from the coordinated actions of all players on the team during the play.

Grunz et al. train a Kohonen self-organizing map on optical data from soccer games. They pick five moments from each play, associate each moment with the closest cluster on the map, and then attempt to classify the resulting length-five vector of clusters as either a "short-game" or "long-game" initiation.

Data

We have been in touch with an NBA team's front office, who have graciously given us their positional tracking data for approximately 400 games of the 2012-2013 season.

The data consists of optical files, which provide an (x, y) coordinate pair for the basketball as well as every player on the court (offense and defense). Each such twelve-tuple is referred to as a *moment*, and the data contains 25 moments per second of play (totaling about 15gb). Additionally, separate metadata files track the occurrence of events, such as dribbles, passes and shots. Because our focus is strictly limited to offensive play-calling, we built a parsing system to attempt to match the events that start and end possessions with the associated optical data. We extract only the offensive player positions and perform a rotation and reflection transformation so that all elements of the dataset are oriented in an identical direction (since in reality teams shoot at two different hoops, this means that in our dataset offensive players are always shooting on a basket on the left-hand side of the screen).

Methodology

Algorithms

Following the prescription of Grunz et al., we felt that a self-organizing map (SOM) would be an appropriate unsuper-

vised algorithm. We relied on the Simon Levy's MATLAB implementation, and chose a 10x10 map as a compromise from the 20x20 grid specified for a larger soccer field in the Grunz et al. paper. Because of the unsupervised nature of our design, we were forced to settle on model parameters by examining the output. To this aim, we built a system that animated the raw optical data into short digital movies so that we could manually inspect the results of our clustering.

We also implemented k-means clustering and, by inspection, it seemed to produce similar centroids to the SOM. As a result, since k-means is a simpler algorithm, it is our preferred clustering algorithm. Our choice of k was 50, which we felt should account for the approximately 15 basic plays that most teams run, as well as accurately capture outliers. We performed sensitivity analysis on the choice of k (see Results).

Finally, we used principal component analysis (PCA) as a validation step on our feature set. While different basketball teams run very different offenses, almost all are broadly composed of some basic basketball strategies, such as setting picks and screens and having fast players make cuts into open space. Our hypothesis was that the principal components of a robust feature set should not vary much by team, since a good feature set would allow the PCA algorithm to break the plays into their underlying basketball components. These components may also correspond to the basic components manually specified by Perse et al.

Challenges

We identify two key challenges in doing unsupervised learning of play-calling on optical data. First, plays develop across time and so the learning algorithm must be able to incorporate information from many static moments. Additionally, the plays each last a variable amount of time, so an exact, one to one comparison of positional data is impossible.

The second major challenge is that keeping player positions in their (x, y) coordinate form forces the ordering of points to be significant. What this means is that a sequence of (x, y) pairs where the order is Player 1, Player 2, ..., Player 5 is only logically comparable to another sequence where the players are ordered identically. This becomes impossible when comparing plays involving many different subsets of players, and in order to compare all permutations, the complexity grows factorially.

Experimental Design

We divide the offensive half of the court into a 25x26 grid (each square represents a two-foot by two-foot square), and if during the course of the play any player or the ball enters a particular grid point, it is marked as active. Thus, each play can be represented as a single observation of the activation of a subset of the court. We will refer to this as the one-court model. This approach remedies both challenges listed above, as the activations cover the entire span of the play, and activations from different players are (at least initially) indistinguishable so there will be a single unique representation of a given play with no ordering issues.

There are, however, two fundamental drawbacks of this approach. This approach is agnostic to explicit player roles within an offense since all activations are treated the same. This trades off information for computational feasibility. To mediate this, we expand our initial model to generate three grids one for the ball, one for shorter players, and one for taller players in subsequent analyses, so that at least some information about player roles is included. We call this the three-court model. The second issue is that flattening all movement across time means that the learning algorithm cannot differentiate between plays that cover the same physical territory in different order. However, from a basketball strategy standpoint it makes intuitive sense, as coaches use a similar, flat diagram structure when designing and communicating plays to the players.

Results

As a sanity check, we used PCA on each team's individual play data using the one-court model to determine whether it is possible to derive relevant information about plays from the features. Figure 1 is one of the top components, and appears to represent a guard's movement from the top of the key to a corner followed by a cut to the basket; this is a common element in plays.

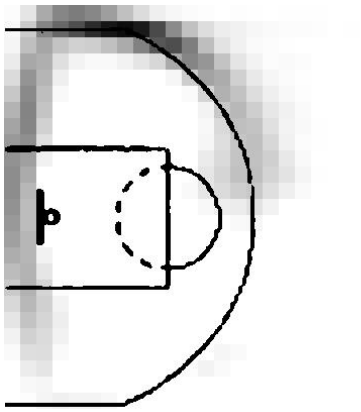

Figure 1: Sample principal component on one court

This indicates that our feature set still preserves a majority of the most important information about plays. We ran the three-court model with similar results. To validate that we are capturing similar information between teams, we performed PCAs for plays of separate teams by splitting up all of the possession data by team. Figure 2 shows the top component for two teams. Despite their very different play styles, these teams share extremely similar principal components. This validates the robustness of the feature set for analysis.

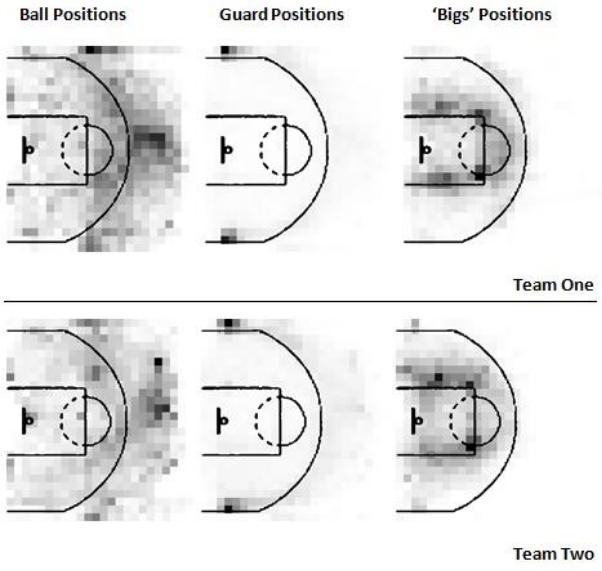

Figure 2: Comparison of the top PCA component of three courts between two teams

As mentioned, Grunz et al. demonstrated that SOM is an appropriate tool to group plays, but SOMs are somewhat expensive to calculate and suggests a relationship between groups based on the shape of the map that we cannot establish. K-means provides a simpler way to group plays and imposes fewer assumptions about the data. We used the following procedure to validate k-means as an appropriate method of grouping plays. An SOM of 10-by-10 neurons and k-means with k = 50 were generated from the data based on the above definitions of features (using the simple one-court model with all players treated identically). A 5-by-5 subsection of the SOM is shown in Figure 3, and two samples of the k-means are shown in Figure 4. The results are similar, which suggests that k-means is an appropriate method of analysis.

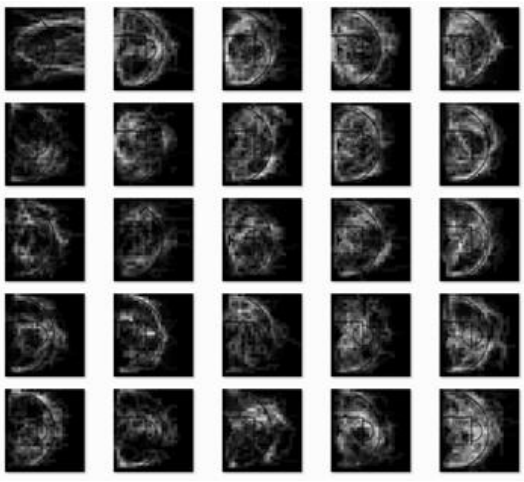

Figure 3: A section of the 10-by-10 SOM

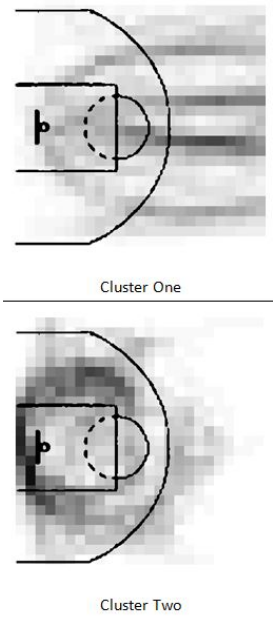

Figure 4: Sample k-means clusters using one grid

In order to determine the best number of clusters, we performed sensitivity analysis by applying k-means with different values of k. An optimal k would have a small number of centroids corresponding to the majority of plays, and the remainder of plays more sparsely corresponding to the remaining centroids. This is because the common plays are heavily utilized by many teams, and plays reserved for high-leverage situations occur less frequently. Figure 5 shows the results of the analysis. It appears that $k = 50$ provides the optimal cluster size; while $k = 200$ actually fits the data better, we believe this is a case of over-fitting because teams do not use that many plays.

Upon estimation of an initial clustering, we noticed that some clusters appeared to correspond to incorrectly classified possessions– in other words, our raw data extraction method was not perfect, and had imprecisely defined some possessions (such as by not terminating when the other team got the ball). We removed plays associated with these centroids from the k=50 model, and performed k-means on the remaining possessions, on a team-by-team basis (Figure 6). For our sample team, this resulted in 89.35% of plays represented by 25 centroids, which is a good representation of frequency of plays in the NBA. Many of the clusters generated using this method resembled plays run by NBA teams. For example, Figure 7 shows one of the clusters that highly resembles the 'horns' play.

One final piece of validation that we ran was to re-estimate the model by dividing each play into thirds time-wise, and generating a nine-court model. We hoped this would allow for closer capture of some of the inter-temporal patterns of movement, and point to even more closely grouped plays. However this seems to have overfit the outliers, as only a couple centroids captured almost all of the total plays. As such, we continue to recommend the three-

court model as defined above.

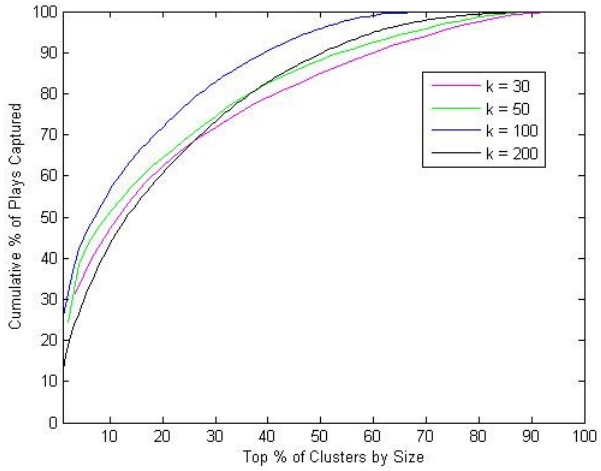

Figure 5: Sensitivity Analysis of different k

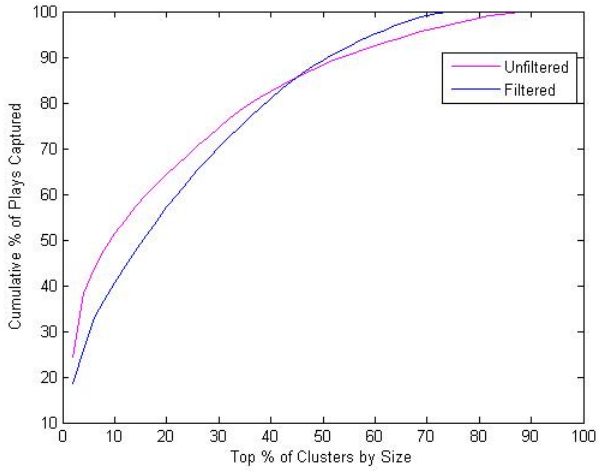

Figure 6: Centroids of Filtered vs Unfiltered Possessions

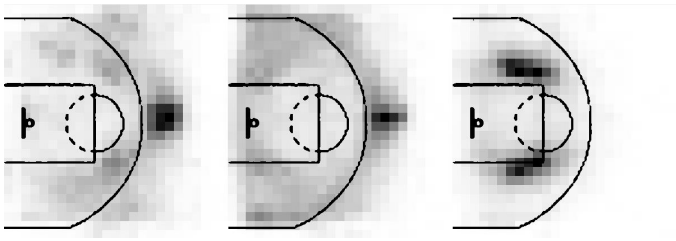

Figure 7: Cluster Resembling 'Horns' Play

Conclusion

The implications of successful classification of basketball play-calling are immense. Teams could divert resources away from the expensive manual capture and identification process, and focus instead on analyzing opponent strategies. On top of being a competitive advantage to the team that implements this first, widespread adoption would hopefully eventually improve the caliber of play as teams focused more on innovative play-calling sequences. We feel strongly that further research in this subject is warranted, and intend to continue our work. We identify several areas in which to improve our approach. First, while initial attempts to incorporate movement through a time vector did not yield improvement in results, we feel this could be improved, perhaps by a more sophisticated segmentation of each play into unevenly sized time blocks. Tackling the factorial growth of permutations of player positions is another important area of work, for with an ability to be agnostic to the initial ordering of player positions, an unsupervised learning algorithm could leverage the gains in greatly decreased feature size as well as additional information of precisely defined player coordinates instead of mapping to a grid.

As it stands, our approach to play-calling would still be valuable to a professional team. By filtering results to only scenarios in which the game was being closely contested, a team could very quickly scan a human-readable diagram of every play an opponent had previously run. Further, outliers from their usual playbook could be marked, so that coaches were aware of special plays that might be run in high-leverage situations, and could forewarn their players to anticipate these. Finally, our data allows teams to estimate the mix of plays that opponents run, and to consider broader strategies of their own to mitigate the advantages that these choices may bring, especially in terms of choosing what type of lineup to expect to use in the game (for instance, playing taller players to prevent the opponent from scoring on plays involving drives to the basket). We hope to continue research toward these goals.

References

1. Hasse, J., and Brefeld, U. (2013). Finding Similar Movements in Positional Data Streams.
2. Perse, M., Kristan, M., Kovacic, S., Vuckovic, G., and Perl, J. (2009). A trajectory-based analysis of coordinated team activity in a basketball game. Computer Vision and Image Understanding, 113(5), 612-621.
3. Grunz, A., Memmert, D., and Perl, J., (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. Human Movement Science 31 334-343.