NBA Players' Performance Prediction in Unsupervised Learning

Yujun Jiang

Machine Learning

Prof. Aaron Hill

April 8, 2020

As a basketball fan and a person who is studying in the data visualization program, I always have a question: how can we use the existing data to predict each player's performance out of the court? I heard Golden State Warriors is the first team which uses data to create the customized training plan and make team decisions based on the predictions. Recently, I recognized many people have invested their time to make athletic performance comparison and prediction by unsupervised learning. Therefore, I think it could be a possible way to let me find an answer for myself, and I will discuss different techniques to achieve my goals.

In unsupervised learning, we have to methods cluster analysis and principle component analysis (PCA). As the traditional way, I would like to find all the related variables such as field goal, assists, blocks, and rebounds to create features. Meanwhile, I can calculate the average value of each feature by using the mean method. For example, I got the average age of an NBA player for 2013-2014 season is 26.5, and I can except the average player to get 516 points, 24 blocks, 39 steals, and 113 assists. Then, I could make 5 clusters (can be more than 5) of players using the machine learning model called K-means to show which players are most similar. To test the model, 80% of the data will be used as training set and the rest of them are test set. At this moment, we will see both Stephen Curry and LeBron James' performance are above the average, but they are playing with the different positions in their teams. The goal is to help team managers and coaches to have a right decision in the daily team training, tactic development, and even for the player trading during each season.

However, the cluster analysis is not a perfect solution in the current NBA league with many changes that have happened. In the past, we can define a player's position in a second because they have some fixed play styles and hot spots. So the cluster analysis could be a reasonable option to evaluate player' performance in the past, it is not enough make a

comparison or prediction with the improvement of athletic ability. Many NBA stars have various and unfixed preference on the court, and that is why the terms like combo and swingman has been redefined. In PCA, it has a technique called future extraction which is to build “new” independent variables as the combination of each the existing independent variables. In this case, I would recommend people to use the cluster analysis as the beginning and add PCA to reduce the dimension of the feature space, it can help us to be more precise in decision making.