NBA Oracle

Matthew Beckler, Hongfei Wang
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
{mbeckler, hongfei}@cmu.edu

Michael Papamichael
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
mpapamic@cs.cmu.edu

1 Introduction

The importance of well-informed decisions regarding player aquisitions [1] and predicting game outcomes in the professional sports business is critical [2]; even moreso in the NBA, which is a multi-billion dollar industry on its own [3]. The goal of this project is to learn, explore, and apply machine learning techniques to an existing dataset of NBA and ABA basketball statistics to: $1)$ Predict the outcome of a game, given the two participating teams, and $2)$ Identify outstanding players based on season and career statistics. In addition to our two original goals, after having extensively examined and worked with our data set and having learned about clustering techniques in class, we are now also interested in applying machine learning clustering techniques to infer player positions.

2 Data Set

The data set used for this project is National Basketball Association (NBA) and American Basketball Association (ABA) statistical data, maintained by *www.databaseBasketball.com*. The data contains player and team statistics from NBA and ABA games throughout the history of these two leagues. Most of the available data is provided for download through the website, but some of the data is not directly downloadable. Some of this extra data includes individual game results, team rosters, and individual game box scores. We created a number of Python scripts to perform *screen scraping* of the website pages, to transform the website-only data in text files usable by our machine learning algorithms.

While plain text files are very flexible and usable with nearly any program or language, with larger and more complicated queries it is often desirable to use a relational database. For this project, we have created a very small script that creates a small, single file SQLite database, and populates it with the downloaded data. Using the expressive nature of SQL allows us to perform large table joins that would be difficult or impossible to duplicate with simple scripts.

With any dataset, it is expected that a certain amount of data conditioning will be required before applying any machine learning algorithms. This dataset is no exception, and we have spent a fair bit of time getting the data prepared. For example, in the `players.csv` file, sometimes the player's college would be listed as "University of California, Los Angeles", where the extra comma would cause problems with CSV parsing utilities. We converted these occurences to conform to the standard CSV format.

We are also interested in performing transformations on the raw dataset to produce better or more expressive data values. Applying transformations to the input data before passing it on to the machine learning algorithms was shown to be more effective than training on the raw data itself. For example, when predicting game outcomes, it was more effective to consider the ratios of certain team statistics instead of the absolute values.

One area in which we are continuing to work is to develop a set of cumulative data from game to game, to give us more accurate and timely data, without "cheating" by using data generated in the

future. An example of this would be predicting the outcome of a game on December 1, 1996; We certainly want to use data from the 1995-1996 season, but we could be more accurate if we included data from the 1996-1997 season for all games occuring before the current game, up through the end of November.

3 Machine Learning Algorithms

3.1 Current Status

To get a better feel for our original dataset, as well as for the additional statistics we obtained through our scripts we initially experimented with very simple classifires. To better understand the importance of each feature to the outcome of a game we tried predicting using only a single feature from our dataset at a time. As expected the most dominant feature in our datasets was the number of wins and losses for each team in the previous season, which already gave an accuracy of 65.9%. We also discovered that there is a strong correlation between the outcome of a game and the number of points received by a team (62.5%), as well as the number of field goals and three-pointers made.

As a next step, to get a better idea of what our target prediction accuracy should be, we examined related previous work for predicting game outcomes, as well as online services that offer betting tips. [4] reports an accuracy of 64.8% when using logistic regression to predict NFL games. [5] used neural networks to predict Soccer matches with accuracy of 65.5%. Finally, [6] reports that experts in the field of professional basketball achieve 71% accuracy. Note though, that the experts have the option to withhold a prediction for hard-to-call games. Even companies in the business of predicting game outcomes only advertise prediction accuracies on the order of 65% [7].

Logistic Regression is a widely used technique for classification used in statistics and machine learning. For the purposes of our project we used logistic regression to predict the outcome of the games for a particular season using team statistics from the previous season. The specific logistic regression implementation is very similar to the one in problem set 2 (part 4). When predicting the outcome of all 1261 games in the 1996-1997 season using 1995-1996 season data, we achieved an overall classification accuracy of 68.1% with 10-fold cross validation (95% confidence interval is ± 0.026).

The weights generated by the algorithm indicated the most influential features for the outcome of a game, namely (in order of importance): pace[1], #wins, #defensive turnovers, #offensive blocks, #three-pts made. The first two features (pace and #wins) were by far the most dominant and this is expected, since pace is a synthetic statistic that incorporates information from many other features and the #wins is expected to be heavily correlated to the outcome of a game.

Linear Regression is one of the fundamental classifers we saw in class. The main linear regression function we tried on our data is similar to the on from problem set 1 (part 4),

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2$$

The accuracy achieved by linear regression is 65.4%. Given the simplicity of this algorithm compared to other classifers, such as neural networks and SVM, the achieved accuracy is quite good.

Support Vector Machines (SVM) are a popular method of supervised learning based on maximizing the classification margin. As part of the initial exploration of our dataset, we experimented with using SVM for game prediction. In these experiments we used the aggregate team statistics from 1995 to predict game outcomes for 1996 season games. Using leave-one-out cross validation for all 1261 games in the 1996-1997 season, we received an overall classification accuracy of 66.9%. It is interesting to note that the number of support vectors was nearly constant across these runs with an average value of 957 support vectors.

3.2 Future Work

Up to this point, we have focused on predicting the outcome of the games. In addition to further work with dataset transformations, we are also interested in trying some recently learned algorithms,

[1] pace is a synthetic statistic, available since the 1973-74 season in the NBA, that incorporates many other statistics and is in many cases used as a metric for ranking teams

such as clustering. Using clustering techniques would allow us to group players into clusters, and perhaps learn what position they play, if they are a real standout player, or possibly reveal some other underlying rules or recurring patterns. A related topic we plan to investigate is outlier detection, which is very useful in recruiting, drafting, and trading decisions.

References

[1] Colet, E. and Parker, J. *Advanced Scout: Data mining and knowledge discovery in NBA data.* Data Mining and Knowledge Discovery, Vol. 1, Num. 1, 1997, pp 121 – 125.

[2] McMurray, S. (1995). *Basketball's new high-tech guru.* U.S. News and World Report, December 11, 1995, pp 79 – 80.

[3] Howard, H. *The Explosion of the Business of Sports in the United States.* Nike Seminar, Spring 1998. http://www.unc.edu/ andrewsr/ints092/howards.html

[4] Babak, Hamadani. *Predicting the outcome of NFL games using machine learning.* Project Report for CS229, Stanford University. http://www.stanford.edu/class/cs229/proj2006/BabakHamadani-PredictingNFLGames.pdf

[5] Balla, Radha-Krishna. *Soccer Match Result Prediction using Neural Networs.* Project report for CS534. http://rk-pvt-projects.googlecode.com/svn/trunk/CS534_ML_SoccerResultPredictor/docs/CS534_ProjectReport.pdf

[6] Orendorff, David, and Johnson, Todd. *First-Order Probabilistic Models for Predicting the Winners of Professional Basketball Games.* Project report. http://www.ics.uci.edu/ dorendor/basket.pdf

[7] AccuScore, *The Leader in Sports Forecasting*, http://tinyurl.com/dhmsts