UNIVERSITY OF NEW SOUTH WALES, SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Predicting Basketball Hall of Fame Membership

COMP9417 - Machine Learning and Data Mining: Project

Chi-Chun Tiao

June 8, 2013

1. INTRODUCTION

Enshrinement into the Naismith Memorial Basketball Hall of Fame is an honour bestowed upon exceptional basketball players, coaches, referees, executives and other major contributors to the sport across the world. As one might expect, the population of inductees is dominated by members in the players category - 158 members in the players category out of a total of 326 members as of the enshrinement class of 2013[1]. Of the 158 players in the Hall of Fame, 103 played in the National Basketball Association (NBA), easily rendering it the league with the most players in the Hall of Fame[1]. As such, we shall limit our discussion to individuals who have played in the NBA. [2]

Unlike other major sport's Hall of Fames where sportswriter voters openly debate their choices, the Naismith Hall of Fame has an opaque and closely held selection process[2]. As a consequence of the screening committee's lack of transparency, the criteria for induction is vague at best and discussion regarding the merits of players is subjective. It would be useful to form an objective basis upon which a player's merits can be analysed and used to predict whether he will be inducted into the Hall of Fame. The goal of this project is to develop such a method using machine learning techniques.

[1] Other inductees of this category include college, international and female players, etc., most of whom have their own shrines already: The College Experience (NCAA), the FIBA Hall (international) and the WBHOF (women's).

[2] Indeed, this inevitably places prejudice against players who had decorated international or college careers, but relatively short or less impressive stints in the NBA, but might nevertheless be considered worthy of Hall of Fame induction.

Currently, some popular methods used include the *Keltner List*, a less objective but systematic method adopted from baseball (as many advanced basketball metrics often are) and also a logistic regression model[4], developed by *Basketball Reference*, a site that provides comprehensive basketball statistics. However, no attempts have been made to develop a machine learning model (to the best of my knowledge). Work has been done to solve the same problem in baseball using a Radial Basis Function Network by Smith, Lloyd and Downey, James (2009)[5], and some of the methods here shall be based upon this.

In this project, I strive to harness the capabilities and advantages of machine learning methods to build a model with better predictive accuracy than that of the aforementioned methods, by experimenting with various different datasets, features and learning algorithms. This report documents the methods employed, work done to collect and preprocess the data, experimentations involved and analysis of the various results.

2. EXPERIMENTATION

2.1. DOMAIN-SPECIFIC FEATURES

In the initial stages of this project, simple basketball statistics were incorporated as features into the training set, alongside other metrics that indicate the relative merit of particular players with respect to other players who played in the league in his time:

- Points Per Game
- Assists Per Game
- Rebounds Per Game
- Career Points
- Career Assists
- Career Rebounds
- Most Valuable Player Awards
- All-Star Game Appearances
- Year Played in the League

Taking into account the fact that a player must be fully retired for five years before being eligible for induction[**?**] and that the inaugural NBA All-Star Game took place in 1951, we used data from players who debuted in the league during or after the 1951-52 season and retired before the 2008-09 season. This not only avoids differences in basketball eras but ensures comprehensive statistics for each player (no missing or incomplete features for any instance) since all the above statistics were being recorded in that time period. Other key statistics such as *steals*, *blocks* and *field goal percentage* were not incorporated for this very reason, as many players either did not have these statistics recorded at all or only had it recorded for some seasons. Other metrics like player's *draft pick*, *weight* and *height* were also considered but these failed to improve performance as shall be discussed later.

In an effort to improve predictive accuracy on this model built on a relatively small number of features, several advanced basketball statistics were considered. Namely, *win shares*, a metric adopted from baseball that is designed to estimate a player's contribution to his team winning, which is made up of two other metrics, *defensive win shares* and *offensive win shares*, each of which considers a players offensive and defensive contributions respectively. Since

it has been said that players are inducted into the Hall of Fame based not only on statistics, but on their defensive prowess, which may not reflect well on a stat sheet, I decided it would be beneficial to incorporate both defensive and offensive win shares. Another statistic considered was the *Player Efficiency Rating (PER)*, a rather arcane metric which summarises a player's performance and contributions into a one number using a very detailed formula. Finally, the statistics *effective field goal percentage (eFG%)* and *True Shooting Percentage (TS%)* were considered, both of which incorporate 3-point and 2-point field goals, the former of which adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal and the latter also considers free throw percentage.

The last metric considered was the number of *NBA Championship Titles* the player has won.

As noted earlier, since many statistics were not being recorded prior to the 1973-74 season, we cannot simply incorporate these advanced basketball metrics into our model as most rely on these missing statistics. Luckily, *Basketball Reference* provides a complete system with estimations to account for missing statistics in particular eras, the formulae for which are quite detailed. The method for calculating win shares and player efficiency rating can be found at `http://www.basketball-reference.com/about/ws.html` and `http://www.basketball-reference.com/about/per.html`.

2.2. Data Collection and Preparation

A considerable challenge faced in this project was the absence of a readily available data repository relevant to the problem, so a substantial amount of work was devoted overcoming this by web data extraction (scraping) and parsing data into a format that could be analysed with machine learning software. The documentation of the data collection process is kept as brief as possible since while part of the overall machine learning and data mining process, it has little to do with machine learning itself.

At first, official NBA stats portal `http://stats.nba.com/` seemed like a promising avenue for data collection. However, not only were there no datasets available for download, the site did not provide any APIs. To avoid resorting to scraping the pages, I inspected the source of the site's pages to find ways to access the backend scripts that drove the site's data and was able to find Javascript sources that provided results in a friendly Javascript Object Notation (JSON) format. Since each source only provided a subset of features for the particular player queried, I was required the cross reference these data sources to create a complete set of features for each player. The data sources and the respective features is provides are described below:

- `http://stats.nba.com/stats/commonallplayers` - *List of all NBA players*
- `http://stats.nba.com/stats/commonplayerinfo/?PlayerID=<player_id>` - *All-Star Appearances (Not available for active players), Position*

- `http://stats.nba.com/feeds/players/profile/<player_id>_Profile.js` - *Draft Pick, Weight, Height*
- `http://stats.nba.com/feeds/players/awards/<player_id>_Award.js` - *Most Valuable Player Awards, Hall of Fame Inductee*
- `http://stats.nba.com/stats/playerprofile` - *Points Per Game, Assists Per Game, Rebounds Per Game, Career Points, Career Assists, Career Rebounds*

At this stage of the data collection process, I hadn't conceived of using the advanced statistics described in the previous section, so these features seemed to be enough to get started on preliminary experimentation. This data source obviously failed to yield the number of championship titles, the aforementioned advanced statistics and the full set of simple statistics that were required to calculate complex metrics like the PER. This posed a problem after no further improvements could be obtained with the original dataset and its limited feature space and an alternative data source was required.

Since Basketball Reference provides all of the aforementioned features, with all of the advanced metrics already calculated (and much more), I finally resorted extracting data from its web pages since there was no API or datasets available for download. The list of players and their respective identifiers were obtained from `http://www.basketball-reference.com/players/<first_letter_last_name>` and a players complete profile could be obtained from `http://www.basketball-reference.com/players/<first_letter_last_name>/<identifier>.html`.

The upside of this was that unlike the official NBA stats portal where it was necessary to piece together data from different sources, it was possible to obtain all the information required from a single page. Using the Python library Beautiful Soup, I was able to extract all the required features for every player. Additional statistics were also obtained in case they were required, such as statistics from Playoff games, All-Star Games, etc. This data source seemed far more comprehensive, and appeared to also include players from the American Basketball Association (ABA), which was later merged with the NBA in 1976. This also meant a larger training set to work with. Furthermore, the Most Valuable Player Award data is expressed in terms of cumulative award shares for the player's career, which is calculated by the number of points a player received for a particular award over the total points of all first-place votes[**?**] - far more useful than merely the raw number of MVP awards in the case of a player such as Jason Kidd, who has never won the award but has been a serious contender in many seasons due to his stellar performance.

Both of the aforementioned datasets have been saved in the JSON format and is included with the submission files. Both datasets can also be parsed into Weka's ARFF or vectorised to work with the Python machine learning library scikit-learn with the provided scripts.

A. APPENDIX

REFERENCES

[1] *"Hall of Famers"* Retrieved 6 June 2013 from Naismith Memorial Basketball Hall of Fame: http://www.hoophall.com/hall-of-famers-index/

[2] Aschburner, Steve. (12 August 2012) *"Hall of Fame selection process leaves much to be desired"*. Retrieved 6 June 2012 from NBA: http://www.nba.com/2010/news/features/steve_aschburner/08/12/hall.process/index.html

[3] Ziller, Tom. (30 March 2010) *"Fans to Vote for Basketball Hall of Fame Inductees"*. Retrieved 1 June 2013 from AOL News: http://nba.fanhouse.com/2010/03/30/fans-to-vote-for-basketball-hall-of-fame-inductees/

[4] *"Hall of Fame Probability"*. Retrieved 1 June 2013 from Basketball Reference: http://www.basketball-reference.com/about/hof_prob.html

[5] Smith, Lloyd and Downey, James (2009) *"Predicting Baseball Hall of Fame Membership using a Radial Basis Function Network,"* Journal of Quantitative Analysis in Sports: Vol. 5: Iss. 1, Article 6.