INTRODUCING MACHINE LEARNING VIA BASEBALL'S HALL OF FAME

David M. Hansen
Department of Electrical Engineering and Computer Science
George Fox University
Newberg, OR 97132
503 554-2709
dhansen@georgefox.edu

ABSTRACT

Machine Learning via Artificial Neural Networks (ANNs) is often introduced in a one-semester course on Artificial Intelligence. Baseball's annual Hall of Fame election provides a simple, tractable, data-rich domain for learning how to use ANNs for predictive analytics. We describe how we use the Fast Artificial Neural Network (FANN) toolkit for a course assignment that predicts which players are likely to be elected to Baseball's Hall of Fame.

INTRODUCTION

Machine Learning via Artificial Neural Networks (ANNs) is often introduced in a one-semester course on Artificial Intelligence [1, 2, 3, 4]. The domain of Major League Baseball and its associated "Hall of Fame" (HOF) provide a simple, yet powerful demonstration of how ANNs can be used for predictive analytics. Our one-semester course in Artificial Intelligence incorporates an assignment that introduces students to the data domain along with a toolkit for building ANNs and tasks them with generating Hall of Fame predictions for current and recently retired baseball players.

THE DOMAIN

Baseball is an ideal domain for a number of reasons. First, baseball has a very long and stable history of data collection with a variety of useful statistics that are publicly available. Second, the data set is significant, but tractable, consisting of data for approximately 30,000 players over the history of the game, ~250 of which have been elected to the HOF. Third, it's generally accepted that a player's election to the HOF is largely tied to their performance so we can expect that a machine learning approach ought to be viable. Finally, students have the opportunity to validate predictions by comparing predictions to the results of each year's election.

Each year a set of recently retired baseball players become eligible for election to the HOF. Sports writers from around the country cast ballots, voting for those players they feel are worthy of the HOF; a player must appear on 75% of the ballots to be elected to the HOF. Eligible players remain on the ballot until 1) they are elected, 2) the time limit expires (currently 10 years) or 3), they fail to appear on at least 5% of the ballots. A

few of the most outstanding players will be elected in their first year of eligibility; most players will either gradually achieve the 75% threshold or briefly peak short of that number and gradually decline until they fall below the 5% threshold or their time-limit expires.

There are, however, a few minor drawbacks to this domain. For one, many students are unfamiliar with baseball and have little understanding of the data. Fortunately one need not understand all the nuances of the domain to work with the data. A second minor drawback is the effect of recent scandals related to the use of "performance enhancing drugs." As players tainted by this scandal become eligible for the HOF their failure to be elected can adversely affect the ability of the ANN to learn that performance is related to electability – especially with the small number of players who are elected. We will discuss how we have chosen to handle these players below.

DATA PREPARATION

The goal of the assignment given to students is relatively simple: given a set of historical data for training, develop an ANN that can be used to predict HOF election for current and recently retired players who are, or will be, eligible for election to the HOF. In order to simplify the assignment for students, we provide them with comma-separated data files that are extracted from a database of baseball statistics.

We use the extensive database curated by Sean Lahman [5] that can be loaded into a DBMS such as MS-Access or MySQL from which data files for student use are extracted via SQL. The first two data files consist of data for "position players" (i.e., non-pitchers); one file holds training data for historical players the other holds data for current players and recently retired players that are eligible for, but not yet elected to, the HOF. The second pair of data files contain data for pitchers. Note that there can be some overlap between the historical data file and the file of current and recent players since recently retired players will appear in both unless they have been elected to the HOF or are no longer eligible to appear on the ballot. For players eligible, but not yet elected to the HOF, the data file includes the highest percentage of HOF ballots they have achieved to date (i.e., $0 \leq n < .75$).

As most of the baseball data is numeric, it requires little or no additional work to prepare it for use with ANNs. However, one important data point for position players is the defensive position played. It is widely believed that some defensive positions have lower offensive expectations for entering the HOF; catchers, for example, are prized for their non-offensive skills. Thus defensive position is important information in training an ANN as position played may make a difference, all other things being equal. Defensive position can be treated as a kind of non-numeric "nominal categorical" data [8]. Although baseball scorekeepers assign numbers to each defensive position (i.e., pitcher=1, catcher=2, etc.) it is generally a mistake to map nominal categorical data onto numeric values because the ANN will try to make sense of the meaningless notion that the position "catcher" is somehow "less than" the position "second base." One way to deal with nominal categorical data is to create a binary value for each of the possible categorical values setting one of the values to 1 and the rest to 0. Thus instead of a single "position" category, n mutually exclusive categories are used, one for each defensive

position. This approach works for categories with small numbers of values (e.g., gender) but can become cumbersome for many-valued categories. However, this approach is quite useful for defensive position since most players play more than one position during their career. So for each player, instead of a mutually exclusive "binary" value, we compute the ratio of games played at each position. This sort of "ratio" data is another common way to transform non-numeric data into a format usable by an ANN [7]. The result is a 27-value numeric vector for each position player and a 17-value vector for each pitcher.

As noted earlier, Major League Baseball has recently been tainted by players suspected of using illegal "performance enhancing drugs." Some of these players have now become eligible for the HOF and sports writers have expressed fan disapproval by casting a disproportionately low number of ballots for these players. Barry Bonds who holds both the single-season and career home run records, for example, appeared on only 35% of the ballots in 2014; under any other circumstances Barry Bonds would certainly have been elected in his first year of eligibility. Due to the rather high number of these outliers and their effect on training, our approach has been to remove these players from the historical data file used for training[*]. We continue to include them in the data file of recent and current players and generate HOF predictions for them as seen in Table 1.

ANN DEVELOPMENT

Students are introduced to the open-source Fast Artificial Neural Network (FANN) toolkit [6, 7] and tasked with training two ANNs using the two historical data files and generating predictions for recent and current players. FANN is a relatively low-level toolkit that provides a C-library API for building and using ANNs. The essential code for constructing, training, and saving an ANN for this assignment amounts to invoking 9 FANN functions in about 15 lines of C-code based on the fully-functional examples provided with FANN. FANN provides a number of parameters that can be used to customize and adjust the training process and we invite students to experiment with these parameters to achieve the highest accuracy.

Training an ANN using the data file of historical players takes mere seconds with FANN; predictions for current and recent players are nearly instantaneous. Students typically collect the output from a FANN-based prediction program and recombine the predicted HOF value with the rest of the data in a spreadsheet for post-prediction examination and validation.

RESULTS

Table 1 shows predicted HOF values for players appearing on the HOF ballot in the spring of 2014; the top three candidates were elected having been present on more than the minimum 75% of ballots cast. These predictions were made via a 3-layer ANN using 27 inputs, one output using FANN's `FANN_SIGMOID` function, and 13 nodes in

[*] We also exclude Pete Rose, banned from baseball for gambling on games.

the "hidden" layer using FANN's `FANN_ELLIOT` function. The network was trained using FANN's `FANN_TRAIN_RPROP` algorithm.

Table 1: 2014 HOF Ballot vs. Predictions

Name	*Year on Ballot*	*2014 Vote %*	*HOF Prediction*
Greg Maddux	**1st**	**97.20%**	**100%**
Tom Glavine	**1st**	**91.90%**	**99%**
Frank Thomas	**1st**	**83.70%**	**100%**
Craig Biggio	2nd	74.80%	100%
Mike Piazza	2nd	62.20%	81%
Jack Morris	15th	61.50%	64%
Jeff Bagwell	4th	54.30%	100%
Tim Raines	7th	46.10%	69%
Roger Clemens*	2nd	35.40%	99%
Barry Bonds*	2nd	34.70%	100%
Lee Smith	12th	29.90%	72%
Curt Schilling	2nd	29.20%	58%
Edgar Martinez	5th	25.20%	49%
Alan Trammell	13th	20.80%	48%
Mike Mussina	1st	20.30%	58%
Jeff Kent	1st	15.20%	34%
Fred McGriff	5th	11.70%	39%
Mark McGwire*	8th	11.00%	95%
Larry Walker	4th	10.20%	33%
Don Mattingly	14th	8.20%	36%
Sammy Sosa*	2nd	7.20%	92%
Rafael Palmeiro*	4th	4.40%	69%
Moises Alou	1st	1.10%	2%
Hideo Nomo	1st	1.10%	0%
Luis Gonzalez	1st	0.90%	14%
Eric Gagne	1st	0.40%	2%
J.T. Snow	1st	0.40%	0%
Kenny Rogers	1st	0.20%	16%
Jacque Jones	1st	0.20%	0%
Armando Benitez	1st	0.20%	0%
Sean Casey	1st	0.00%	5%
Todd Jones	1st	0.00%	0%
Ray Durham	1st	0.00%	1%
Mike Timlin	1st	0.00%	0%
Paul Lo Duca	1st	0.00%	1%
Richie Sexson	1st	0.00%	0%

* Player suspected of using "performance enhancing drugs"

If these predictions are to be believed, players such as Craig Biggio and Mike Piazza are likely to be elected at some point in the future as our model's prediction is the maximum expected for that player during their time on the ballot.

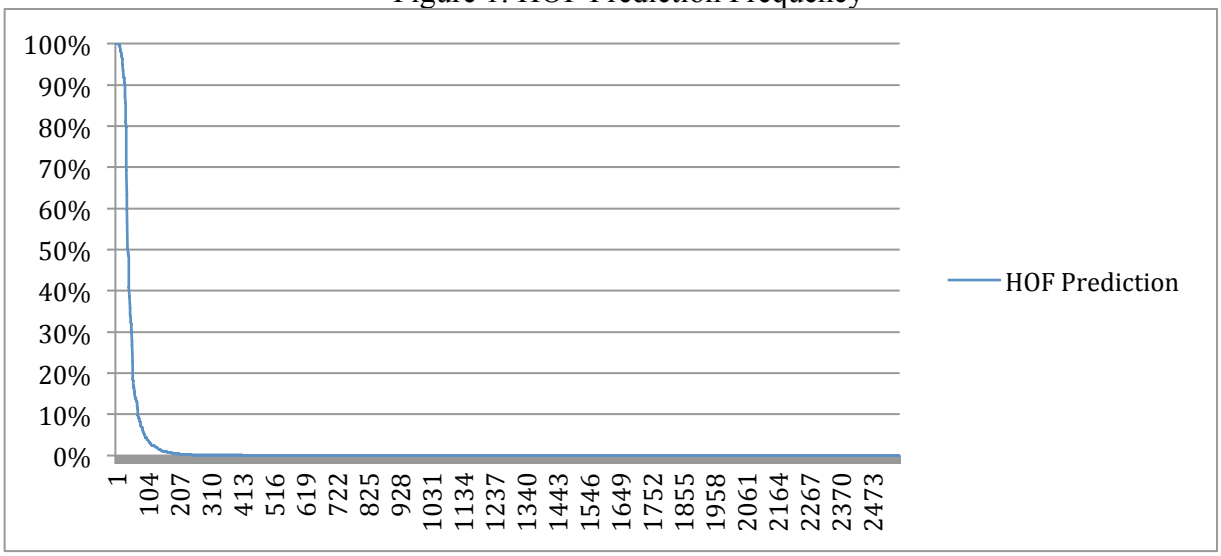
Table 2 presents a subset of the predictions for current and recently retired players who are not yet eligible to appear on the HOF ballot. Of 2565 position players evaluated, the network predicts that the 28 players shown in Table 2 are likely to rise above the 75% threshold. Of course some of these players are likely to fall short as the prediction is based on a very few games (e.g., Dusty Wathan) or the player is suspected of using "performance enhancing drugs" (e.g., Alex Rodriguez). Those caveats aside, most any knowledgeable baseball fan would likely agree with most of the names presented in Table 2.

Table 2: HOF Predictions for Current/Recent Players

Name	***HOF Prediction***	***Games***
Alex Rodriguez	100%	2568
Derek Jeter	100%	2602
Ken Griffey	100%	2671
Miguel Cabrera	100%	1660
Ichiro Suzuki	100%	2061
Albert Pujols	100%	1958
Vladimir Guerrero	100%	2147
Ryan Braun	100%	944
Ivan Rodriguez	100%	2543
Alfonso Soriano	100%	1908
Gary Sheffield	100%	2576
Dusty Wathan	99%	3
Josh Hamilton	98%	888
Chipper Jones	98%	2499
Mike Trout	98%	336
Andrew McCutchen	98%	734
Carlos Beltran	97%	2064
Jose Reyes	97%	1303
Jimmy Rollins	96%	1952
Michael Young	95%	1970
Manny Ramirez	93%	2302
Matt Holliday	93%	1434
Rafael Ortega	92%	2
Lance Berkman	91%	1879
Bryce Harper	90%	257
Rick Short	87%	11
Larry Gonzales	86%	2
Todd Helton	80%	2247

Figure 1 provides further evidence of the effectiveness of ANNs for HOF prediction. Figure 1 shows the distribution of HOF predictions for all recent and current players (i.e., results from which Table 2 is extracted). The ANN very effectively discriminates between the vast majority of players and those few that are worthy of consideration for the Hall of Fame.

Figure 1: HOF Prediction Frequency



CONCLUSION

Major League Baseball and its "Hall of Fame" provide a useful domain for introducing students to Machine Learning via ANNs within the scope of a one-semester course in Artificial Intelligence. Election to the HOF is clearly a function of the player's career that is well characterized by available data and ANNs can quickly and easily be trained to recognize HOF candidates.

While we have chosen to use the FANN toolkit, a number of toolkits and software systems are available for ANN development including popular commercial tools such as Matlab, SAS, SPSS, etc. Among open-source Machine Learning toolkits, the Java-based Weka toolkit (http://www.cs.waikato.ac.nz/ml/weka/) appears to offer help in automating portions of the data transformation task above [8]. More recently, web-services such as the Google Cloud Prediction API (https://cloud.google.com/products/prediction-api/) and BigML (https://bigml.com/) are beginning to offer online tools and APIs to facilitate predictive analytics.

REFERENCES

[1] Poole, D., Mackworth, A., *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, 2010.

[2] Russell, S., Norvig, P., *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[3] Coppin, B., *Artificial Intelligence Illuminated*. Jones and Bartlett, 2004.

[4] Luger, G., *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison Wesley, 2009.

[5] http://www.seanlahman.com/baseball-archive/statistics/ accessed February 2013.

[6] S. Nissen, *Implementation of a Fast Artificial Neural Network Library (FANN)*, Department of Computer Science University of Copenhagen (DIKU), Tech. Rep., 2003.

[7] Fast Artificial Neural Network Library (FANN) from http://leenissen.dk/fann/wp/ accessed February 2013.

[8] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2011.