Forecasting Baseball

Clint Riley
clintr@stanford.edu

December 14, 2012

Abstract

Forecasts for the outcome of sporting events are coveted by nearly everyone in the sporting world. In this paper, a number of machine learning algorithms for predicting the outcome of baseball games are explored, using both classification and regression approaches.

1 Introduction

Baseball is America‘s national pastime and has become increasingly popular around the world in recent decades. Wherever baseball goes, the boxscore, and massive amounts of data in the form of statistics follow. Because of the sheer volume and detail of available data for the game (especially in Major League Baseball, where accuracy of scorekeeping approaches 100%), it lends itself very well to be analyzed and forecast using statistics and machine learning algorithms. The effectiveness of various learning algorithms in predicting the outcomes of games from previous data is explored below.

2 Data Collection

There are a multitude of websites and many dozens of books offering baseball statistics in varying granularities (see references for a small sample). Most of these sources, however, do not present their data in a way that is amenable for automated processing.

Retrosheet.org is the source for the data that was used in this project. Retrosheet presents their data in files offering pitch by pitch granularity with files covering as far back as the 1940s (although older data has a tendency to be less detailed and less accurate)[1]. Additionally, Retrosheet provides tools that handle their specific data format and output CSV files according to various parameters. Using these files, along with a few dozen SQL queries. it was possible to create a SQL database containing all of the relevant information and to produce a huge amount of training data, with nearly every desired feature.

Unfortunately, large amounts data brings its own baggage, and calculating the desired features turned out to be non-trivial from a database with many millions of rows, so it wasn‘t possible (in this short timeframe) to exhaust the amount of training data that is theoretically derivable from the database.

3 Features

3.1 Sabermetrics

Sabermetrics is a term that is derived from the acronym of the Society for American Baseball Research (SABR). Bill James (now employed by the Boston Red Sox), one of the pioneers of the area, invented the term. The mantra of the Sabermetrics community is the search for objective knowledge about baseball, a concept that dovetails well with this paper.

Baseball statistics and boxscores have been around nearly as long as the game, and in massive quantities. A number of the more prevalent statistics, however, have significant problems. Consider wins for a pitcher, for instance. The official rule for determining a winning pitcher is several paragraphs long, but the major flaw is that it depends as much on the teams offense (which the pitcher has zero or minimal influence over), as it does on the pitchers performance. A hurler could pitch nearly flawlessly and not be credited with a win (for example, consider Ken Johnson, who, remarkably, didnt allow a single hit during an outing in 1964, yet was charged with a loss). Conversely, a pitcher could allow a dozen runs and still pick up a win if his teams offense is even more explosive. With these caveats in mind, the features used in this paper were chosen with the goal of accurately reflecting a players impact on runs scored, the ultimate predictor of wins and losses [6] [12]. These statistics were tested and shown to have a nonzero correlation with the number of runs scored [2] [4].

Batting Statistics

1. OBP - On Base percentage
2. SLG - Slugging percentage
3. OPS+ - League adjusted on base percentage plus slugging percentage.
4. RC/G - Runs created per game
5. ISO - Isolated power
6. SB/CS - Stolen base to caught stealing ratio
7. K/G - Strikeouts per game
8. K/BB - Strikeouts to walks ratio

Pitching Statistics

1. ERA+ - League adjusted earned run average
2. WHIP - Walks plus hits per inning pitched
3. K/G - Strikeouts per game
4. BB/G - Walks per game
5. K/BB - Strikeouts to walks ratio
6. The above batting statistics, for batters facing the pitcher.

Fielding Statistics

1. FLDP - Fielding percentage

In an effort to reflect the aging curve, which indicates a player will ramp up to his potential, peak, and then decline towards the end of his career [7], these statistics were taken over differing time periods. The short term period covers the previous three months, the medium term covers the previous two seasons, and the long term covers the player‘s entire career. A period shorter than three months is essentially meaningless due to the random nature of the game and since momentum in sports has been shown to be largely a myth, outside of certain niche sports [10].

Additionally, two ”plus” statistics were used, OPS+ and ERA+, which aim to isolate the player‘s performance from the stadiums he has played in. This is known as adjusting for the park factor. Two of the more extreme parks are home to teams in the NL West division: Coors Field (home of the Rockies) and AT&T Park (home of the Giants). Simplistically, a park factor is a number that is multiplied with statistics to normalize them (Coors Field, a very offense friendly park, has a park factor of 1.579, while AT&T has a park factor of just .737, making it a friendly confine for the Giants‘ pitching staff) [6].

Table 1: Logistic Regression Results

Training Examples	Training Error	Test Error
1000	.401	.4800
2000	.432	.4788
3000	.456	.4650
4000	.460	.4625

Table 2: Adaptive BoostingResults

Tr. Examples	Weak Classifiers	Test Error
1000	50	.4551
	100	.4487
	200	.4487
2000	50	.4487
	100	.4375
	200	.4338
3000	50	.4363
	100	.4225
	200	.4200
4000	50	.4363
	100	.4225
	200	.4263

4 Learning Algorithms

The most frequently used approach to predicting the outcomes of sporting events is to use a straightforward (or not so straightforward) binary classification. This works reasonably well, but doesnt capture the fashion in which games unfold (i.e. its not possible to tell what the score of the game was from a win/lose label). To tackle this issue, a regression approach was pursued, in addition to classification. The idea of the regression approach is to forecast how many runs a given team will score, compare that to the forecast for the opponent, and decide from these comparisons who is more likely to win.

4.1 Classification

Logistic Regression

As a first step towards experimenting with classification algorithms, plain old logistic regression was run with varying numbers of training examples. As illustrated in Table 1, the results were less than spectacular.

It should be noted, however, that the fully fleshed out feature set was not tested with logistic regression, so it is possible that the results would be improved. That being said, however, there is little reason to believe that plain logistic regression would perform better than the more advanced classification methods.

Adaptive Boosting and Decision Trees

Most of the effort in the classification realm was spent on the AdaBoost algorithm. To review, the basic idea is to leverage a large number of weak classifiers, and weight them appropriately to form one strong classifier [5]. The weak classifiers used in the following experiments were decision trees.

As table 2 indicates, the best result produced from AdaBoost is 58% accuracy, yielded by 200 weak classifiers (decision trees). No improvement was seen moving to 4000 training examples from 3000, but it is likely further gains would be possible with a higher number of training examples, given how noisy baseball data can be.

4.2 Regression

Linear Regression

As another early attempt, simple linear regression was performed, but resulted in unreasonably high error, yielding mean squared error on the run totals of greater than 5. This was abandoned relatively

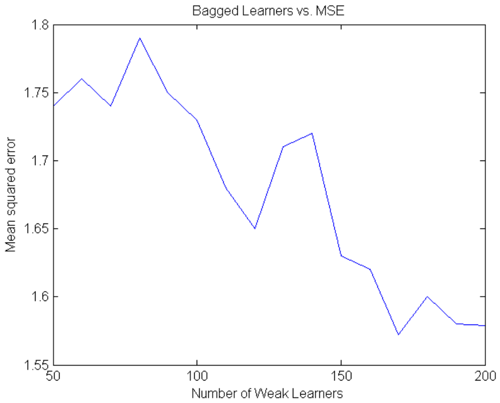

Figure 1: Bagged Regression Trees - MSE

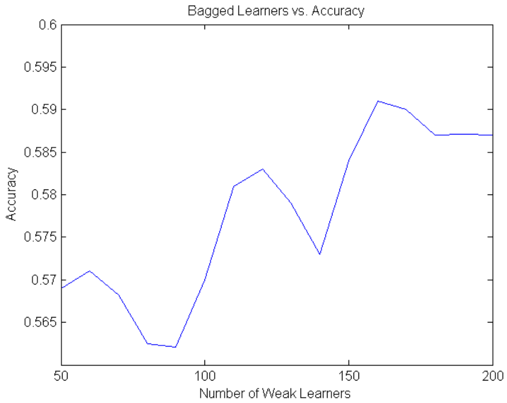

Figure 2: Bagged Regression Trees - ”Classification” Accuracy

quickly.

Bagging and Regression Trees

Bagged regression trees yielded the best result of the considered algorithms for this particular problem and feature set (yielding an accuracy of 59.25%). Like boosting, bagging is an ensemble learning method that utilizes a large number of weak learners. Unlike boosting, however, each one is weighted equally, and each tree is trained on some set of examples chosen randomly with replacement from the entire training set. The predicted result is the mean of the weak learner outcomes.

There are two ways to evaluate the performance of bagged regression trees on the problem at hand. The first is to compare how significantly the predicted run value differs from the actual runs scored. This is reflected in figure 1, with the number of weak learners (regression trees) on the x-axis and the mean squared error on the y-axis.

The second way to evaluate the accuracy is to use the predicted run totals to determine which team would win the game. These results appear in figure 2, with the number of learners again on the x-axis and the accuracy on the y-axis.

These values were calculated using an external test set of 800 examples.

Quantities of learners greater than 200 were briefly tested, but resulted in higher test error due to overfitting.

5 Future Work

5.1 Additional Features and Data

MLB Gameday Data

One thing that is surely worth exploring is the utilization of Major League Baseballs (MLB) gameday data. This contains a truly amazing amount of detail about every single pitch of every single game. MLB has installed specialized cameras in the stadiums designed to track the speed, trajectory, and break of each pitch. From this, it could be possible to find tendencies in pitchers and hitters and leverage that to predict performance.

Fielding and Defensive Statistics

The utilization of fielding and defensive data was notably deficient in this project. Fielding percentage alone does not sufficiently reflect how effective a defensive player or team is. Consider, for example, a player with extremely limited range (that is, he‘s unable to reach balls more than one foot from where he began the play). He may be able to handle these very well and as a result have a high fielding percentage, but clearly he is a much less effective fielder (in preventing runners from reaching base and thus scoring runs) than someone with a larger range who is marginally less sure handed.

Incorporation of External Features

Data outside of baseball statistics could be incorporated into the algorithm. An example would be Twitter sentiment, which has been shown to be useful in some cases. Its unclear if there would be sufficient data from Twitter to produce a meaningful signal for a regular season MLB game, however.

5.2 Other Algorithms

Gradient Boosting

This is another variant of ensemble learning for regression problems that has shown promise for various applications.

References

[1] Adler, Joseph, *Baseball Hacks: Tips & Tools for Analyzing and Winning with Statistics*. 2006: O‘Reilly.

[2] Albert, Jim and Bennett, Jay, *Curve Ball* 2001: Springer.

[3] Bishop, Christopher M., *Pattern Recognition and Machine Learning* 2007: Springer.

[4] Bradbury, J.C. *The Baseball Economist* 2008: Plume.

[5] Freund, Y. and Schapire, Robert E., ”A decision-theoretic generalization of on-line learning and an application to boosting”. *Journal of Computer and System Sciences*. 1997.

[6] Keri, Jonah, *Baseball Between the Numbers* 2007: Basic Books.

[7] Lichtman, Mitchel, *“How do baseball players age?* “ Hardball Times. http://www.hardballtimes.com/main/article/how-do-baseball-players-age-part-1. 2009.

[8] Marsland, Stephen, *Machine Learning: An Algorithmic Perspective* 2009: Chapman and Hall.

[9] Mohri, Rostamizadeh, and Talwalkar, *Foundations of Machine Learning* 2012: The MIT Press.

[10] Moscowitz, Tobias and Wertheim, L. Jon *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won* 2012: Three Rivers Press.

[11] Ng, Andrew, CS229 (Autumn 2012) Stanford Machine Learning lecture notes. http://cs229.stanford.edu/materials.html

[12] Schwarz, Alan *The Numbers Game: Baseball‘s Lifelong Fascination with Statistics* 2005: St. Martin‘s Griffin.

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