Ler e entender esse artigo!!!

DATA MINING IN SPORTS: PREDICTING CY YOUNG AWARD WINNERS*

Lloyd Smith, Bret Lipscomb, and Adam Simkins
Department of Computer Science
Missouri State University
Springfield, Missouri

ABSTRACT

A Bayesian classifier was created to predict Cy Young Award winners in American baseball. The model was compared against two statistical models designed to perform the same task. Over the years from 1967 through 2006, the accuracy of the Bayesian classifier was similar to that of the other two models—when restricted to starting pitchers, all three were more than 80% correct. Accuracy was lower for all three models when relief pitchers were included in the data.

INTRODUCTION

Since 1967, the Cy Young Award has been presented annually to the best pitcher in each of the two leagues of Major League Baseball. Each year, as the season progresses, baseball fans and sportswriters take great interest in predicting the winners of the award. Because the award is based largely on statistics compiled over the course of the baseball season, predicting the winners is a logical application for statistical models and data mining. Because the winners are elected (by members of the Baseball Writers Association of America), however, the criteria are not static and are not always clear; therefore, predicting winners is a challenge.

Perhaps the earliest attempt to predict Cy Young award winners using a statistical model was by Bill James [3], now a player development consultant for the Boston Red Sox. James' model uses a weighted average of wins (for readers unfamiliar with baseball terminology, definitions may be found online in the Wikipedia [5]), losses, saves, strikeouts, shutouts, and "runs saved" (a measure of how

many fewer runs a pitcher gave up compared to a pitcher who gave up five earned runs per nine innings pitched); in addition, the model adds a bonus for pitchers who played on a division winning team. Sparks and Abrahamson developed a similar model [4], using a weighted average of wins, losses, strikeouts, earned run average (ERA), and team winning percentage.

We chose to attack the problem using a data mining approach—rather than develop a static statistical model, we wanted to see if a machine learning algorithm, by examining Cy Young Award winners and their statistics, could learn to reliably predict award winners. If successful, this approach has the advantage of being able to easily incorporate new data each year. If voters are consistent, the model should be able to increase in accuracy over time.

In the remainder of this paper, we introduce basic data mining concepts and the algorithm used in this project, describe experiments we performed in testing the data mining algorithms, and discuss the results of those experiments. Finally, we present our conclusions and suggestions for further work.

DATA MINING

In general, data mining is concerned with discovering structural patterns in data [6]. The learning may fall into several different categories, but we are concerned with algorithms that perform classification—given statistics generated by pitchers over the course of a baseball season, which pitcher is most likely to be classified as a Cy Young Award winner?

The data mining algorithm we used is a naïve Bayes classifier, implemented in the Waikato Environment for Knowledge Analysis (WEKA) [6], a data mining workbench that enables experimentation with a number of data mining methods.

Bayesian Classification

Bayesian classifiers use Bayes' Theorem to perform probabilistic classification. Bayes' Theorem is defined by Formula 1, where P(H|E) is the probability that some hypothesis H is true given that evidence E has been observed, P(H) is the probability that the hypothesis is true without regard to the evidence (the *prior* probability), and P(E) is the prior probability that the evidence will be observed.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$ Formula 1. Bayes' Theorem.

Probabilities are determined by examining a set of training data—statistics from award-winning pitchers determine the probabilities for the "Yes" hypothesis (the hypothesis that the pitcher will win the award); statistics for non-winning pitchers determine the probabilities for the "No" hypothesis (that the pitcher will not win the award). The formula is applied to each pitcher in the test set for both hypotheses—Yes and No; the evidence is the set of statistics for that pitcher. The pitcher with the highest probability for Yes is classified as the winner. In practice, we need only compute the numerator of the formula because the denominator—P(E)—is the same

for both Yes and No for a given pitcher, and may be ignored. The prior probabilities, P(H), are the same for each pitcher; because we are keeping 10 pitchers for each year and league (as discussed below), and because one pitcher in each league wins the award each year, the prior probability of Yes for each pitcher is 1/10, and the prior probability of No is 9/10.

The naïve Bayes classifier is so called because it naively assumes that each attribute is independent of all others and each is equally important. This is often not true, and is certainly not true in our data set. For example, while it is not always the case, a pitcher with a lower ERA will likely win more games and lose fewer games than other pitchers—so ERA, wins, and losses are clearly not independent. It's also true that a pitcher who strikes out a large number of batters is likely to have a lower ERA than other pitchers, so all four attributes are likely to be dependent, to some degree, on the others. Nor are the attributes equally important. Intuitively, it seems likely that sportswriters weight wins most heavily, followed by ERA; both the James and Sparks and Abrahamson models weight attributes in just that way. Still, naïve Bayes is effective in many applications.

EXPERIMENTS

This section describes experiments designed to test the ability of the data mining algorithm to correctly predict the Cy Young Award winners in the years from 1967 to 2006.

Data

We downloaded statistics for the baseball seasons from 1967 through 2005 from the Lahman Baseball Database [1], which holds data for every player and manager that has ever appeared in a major league game. Data for the 2006 season was downloaded from ESPN [2]. We define starting pitchers as those who made more than half their game appearances in a starting role; relief pitchers are those who made half or more of their appearances in relief. For experiments with starting pitchers (described below), we used the top 10 pitchers, in terms of wins, for each league in each year. When dealing with relief pitchers, we used the top 10 in saves.

Our statistics were comprised of the basic attributes we believe sportswriters use in award voting: wins, losses, strikeouts, and earned run average (ERA); when dealing with relief pitchers, we include saves. For our classifier, we normalized the data by converting the numeric attributes to z scores. In converting wins to z scores, for example, we convert each pitcher's number of wins to its number of standard deviations above or below the average number of wins for that league and year.

In statistical models such as those of James and Sparks and Abrahamson, attributes are weighted by their relative importance. Naïve Bayes, however, assumes equal weights for all attributes. In order to weight wins and ERA more strongly than other attributes, we include wins three times, and ERA twice, in the data set. This violates the Bayesian assumption of independence and equal weighting of attributes,

but it does have the desired effect of weighting the repeated attributes more strongly than the others.

The Problem of Relief Pitchers

Most pitchers who win the Cy Young Award are starting pitchers. Since 1967, however, nine relief pitchers have won the award. Over this time, the roles of relief pitchers have changed significantly [3]. In 1974, when Mike Marshall became the first relief pitcher to win the award, a reliever entered the game when the starter tired or became ineffective, and he was expected to finish the game, if possible. The result was that relief pitchers amassed a relatively high number of wins and losses, and a relatively low number of saves. Marshall, for example, pitched 208 1/3 innings and won the award with a record of 15 wins, 12 losses, and 21 saves; three years later, Sparky Lyle pitched 137 innings and won with 13 wins, 5 losses, and 26 saves. Since then, relief pitchers have become more specialized. Wikipedia [5] lists five different kinds of relief pitcher, and the stars among them are the *closers*—pitchers who enter the game almost exclusively when their team is winning by a small margin and pitch only the last one or two innings. The most recent relief pitchers to win the Cy Young Award are Dennis Eckersley, who pitched 80 innings in 1992, winning 7, losing 1, and saving 51 games, and Eric Gagne, who pitched 82 1/3 innings in 2003, winning 2, losing 3, and saving 55 games.

Given enough consistent data, a data mining algorithm can learn to accommodate relief pitchers. However, with the changing roles of relief pitchers and only nine winners, it is a very difficult task. Sparks and Abrahamson [4] deal with this problem by ignoring relief pitchers—they consider only starting pitchers in their model. James incorporates relief pitchers, but admits that his model does not deal with them very well [3].

Our first experiment follows the example of Sparks and Abrahamson, and considers only starting pitchers—we eliminate years in which relief pitchers won the award. Our second experiment considers only relief pitchers—we test only over years in which relief pitchers won the award, and use only relief pitchers in the tests—this enables us to determine whether we can have any hope that our data mining algorithm is able to deal with relief pitchers. Finally, our third experiment combines starters and relievers and attempts to predict the award winners regardless of which type of pitcher won.

Predicting Cy Young Award Winners—Starting Pitchers

In order to test the data mining algorithm, we created train and test sets for each year and league. Train sets include the top 10 pitchers, in number of wins, for each year except for the year and league being tested; test sets include the top 10 winning pitchers only for the year and league being tested. The test set for the American League in 1999, for example, includes the top 10 winning pitchers in the American League in 1999; the corresponding train set includes the top 10 winning pitchers in the American League for each year from 1967 through 1998, and for 2000 through 2005.

This means the tests were carried out in a round-robin process in which test data was never used in training the model used to make a prediction on that set of data.

Eliminating the nine relief pitchers who won the award between 1967 and 2006 leaves 71 classification tasks. Table 1 shows the results of the tests with starters. For comparison, results from the James and Sparks and Abrahamson models are included.

	Total	N correct	% correct
Naïve Bayes	71	60	84.5
James	71	62	87.3
Sparks-Abr	71	61	85.9

Table 1. Prediction of Cy Young Award winners, 1967-2006—starting pitchers only.

The James and Sparks and Abrahamson models have an inherent advantage because those models were built using some or most of the data. Sparks and Abrahamson optimized their model on data from 1992 through 2002; James doesn't report precisely which data he used, but appears to have optimized his model on all data through 2002. Those models, then, were tested over at least some of the data on which they were trained, while the Bayesian model was trained and tested on separate data. At any rate, it is clear that all three models have similar accuracy when dealing with starting pitchers.

Testing Over Relief Pitchers

There were nine years, between 1967 and 2006, in which relief pitchers won the Cy Young Award. Table 2 shows the results of testing only on those years, and using only relief pitchers in the test. For both the naïve Bayes classifier and the Sparks and Abrahamson model, saves were substituted for wins; James uses both wins and saves, so no change was necessary in his model. For this experiment, we used the top 10 relief pitchers, in terms of saves, in each league for each year tested.

	Total	N correct	% correct
Naïve Bayes	9	9	100
James	9	7	77.8
Sparks-Abr	9	7	77.8

Table 2. Prediction of Cy Young Award winners, 1967-2006—relief pitchers only.

The Bayesian classifier was most successful in learning to predict award-winning relief pitchers. These results show the advantage of the learning algorithm over the static models—the data mining model is able to learn how important saves are for relief pitchers. Interestingly, the other two made different errors; the James model weights wins too heavily when dealing with only relief pitchers, while the Sparks and Abrahamson model erred in two years when the reliever with the most saves was not the winner.

Experiments Incorporating Starters and Relievers

Finally, we carried out experiments incorporating both starters and relievers for all years from 1967 through 2006. For these experiments, naïve Bayes treated wins and losses as missing attributes for relief pitchers, and saves as a missing attribute for starting pitchers. The Sparks and Abrahamson model was restored to using wins, rather than saves—in effect, this meant (as intended by its designers) that the model did not consider relief pitchers. Table 3 shows the results of these tests.

	Total	N correct	% correct	Starters correct	Relievers correct
Naïve Bayes	80	56	70.0	52/71	4/9
James	80	59	73.6	56/71	3/9
Sparks-Abr	80	61	76.3	61/71	0/9

Table 3. Prediction of Cy Young Award winners, 1967-2006—all pitchers.

It appears that the best approach is that of Sparks and Abrahamson—because their model ignores relief pitchers, it accurately predicts the same number of winners that it does when dealing only with starters. The other two models correctly predict—in total—fewer winners than when dealing only with starters, and perform poorly on the relievers. The most recent award in the National League illustrates the difficulty of the problem—six pitchers led the league in wins with 16, an unusually low number for the league leader. When restricted to starting pitchers, all three models correctly chose Brandon Webb by a small margin. When including relief pitchers, however, our model chose Trevor Hoffman, who finished second in the voting with an 0-2 record and a league-leading 46 saves. James' model selected Billy Wagner, who had a 3-2 record with 40 saves. James says, "Whatever it is that causes a relief pitcher to ring bells with the voters, the formula doesn't reliably quantify it" [3]. We can extend that to say that it's possible to deal with starting pitchers or relief pitchers, but difficult to deal with both together.

CONCLUSION

The goal of this project was to develop a Bayesian classifier to predict Cy Young Award winners in American major league baseball. The accuracy of the classifier was comparable to that of two statistical models described in the literature [3, 4]. All three of the models are effective when dealing with starting pitchers or relief pitchers, but less effective when dealing with both together. The advantage of the Bayesian classifier, over the other models, is that it is a learning algorithm—it can easily modify itself when incorporating new data in future years.

This method can be extended to other sports awards. The primary requirement is a set of well-defined statistics that are likely to identify the top candidates. The most valuable player (MVP) awards in baseball and basketball, for example, should be amenable to a data mining approach. The MVP award in football is more problematic because of the different nature of the statistics compiled by quarterbacks, running backs, and receivers, all of whom compete for the award.

Possibilidade de continuação do projeto em outros espotes.

Finally, the problems studied and described in this paper could form the basis for lab exercises in statistical modeling or data mining in a CS0 class; while such labs will not capture the interest of all students, they are likely to be of intense interest to some.

REFERENCES

[1] Baseball Archive, Lahman Baseball Database, Version 5.3, www.baseball1.com, retrieved January 18, 2006.

[2] ESPN MLB Sports Index, http://sports.espn.go.com/mlb/statistics, retrieved October 2, 2006.

[3] James, B., E = M Cy squared, in *The Neyer/James Guide to Pitchers,* James, B. and Neyer, R., New York: Simon and Schuster, 467-471, 2004.

[4] Sparks, R. and Abrahamson, D., A mathematical model to predict award winners, *Math Horizons,* April 2005, 5-13.

[5] Wikipedia, www.wikipedia.org, downloaded October 22, 2006.

[6] Witten, I. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd ed., San Francisco, CA: Morgan Kaufmann, 2005.