

Beyond “Playing the Percentages”: Application of Collaborative Filtering for Predicting Baseball Matchups

Joseph C. Bonneau

December 6, 2006

Abstract

Predicting the effectiveness of specific batters against specific batters is a key element of baseball strategy. Evidence suggests that many pairs of professional batters and pitchers who have faced one another many times produce significantly different results than would be predicted with a naive averaging of the batter’s and pitcher’s overall statistics. Yet, for most possible matchups there does not exist enough data to reliably predict from history how a given batter will fare against a pitcher by simply looking at past results. This paper introduces a new approach, based on collaborative-filtering techniques developed in recommendation systems, which can predict the results of batter-pitcher matchups with greater accuracy than either of these simple methods.

1. Introduction

The game of baseball is largely won and lost in the matchup between batters and pitchers in a series of “plate appearances” or “at-bats”. While no two plate-appearances are truly the same, baseball is represented very well by grouping all possible outcomes into a finite number of categories, such as single, home-run, strikeout, double play, etc. Despite the large number of outcomes, the game is reasonably well modeled by collapsing all possible outcomes of a plate appearance into two, namely, all outcomes which result in the batter being called out and all outcomes which involve the batter safely reaching base. This reduction of outcomes leads to the utility of *on-base percentage* (OBP), defined as:

$$OBP = \Theta = \frac{\#successful\ plate\ appearances}{\#total\ plate\ appearances}$$

Roughly speaking, Θ_b for some batter b represents that batter’s success rate in reaching base and not being called “out”. In general, a batter’s effectiveness is indicated by Θ_b . This ignores many important facets of the sport, such as the frequency at which home runs are hit and the frequency at which a batter strikes out, but for this paper will be used as the primary measure of a batter’s effectiveness¹.

¹Historically, *batting average* is much more commonly used for this

Similarly, for a pitcher p , we will use the value Θ_p^{-1} is that pitcher’s *on-base percentage against* (OBPA), which indicates how successful opposing hitters are at reaching base and can be used as a general indication of the effectiveness of the pitcher (a lower value being better). We will also extensively consider the *matchup on-base percentage* $\Theta_{b,p}$, which represents the success rate of a specific batter b against a specific pitcher p .

A large amount of research by both amateur baseball fans and professional analysts has gone into measuring how well statistics like OBP, OBPA, and dozens of others reflect the value of batters and pitchers in helping their teams win games. This type of analysis is useful for overall comparison of players when assembling a roster. During games, however, a roster is fixed, and a manager is forced to use the players at hand.

Managers do have some ability during games to choose which batters from their team will face pitchers from the opposing team, and conversely which pitchers from their team will face opposing batters. For these micro-decisions, it is important to accurately predict specific batter-pitcher matchups, as opposed to the overall abilities of a batter and a pitcher.

For example, it is often the case late in games that a manager has the option of bringing in pinch-hitter, that is, a manager must choose between sending batter b_1 or b_2 to the plate against the opposing pitcher p . A rational manager should choose based on the larger value of $\Theta_{b_1,p}$ and $\Theta_{b_2,p}$ to maximize his teams chances of success. However, estimating accurate values of $\Theta_{b,p}$ is a difficult problem, since in many cases there will be have been very few prior matchups between b and p to infer $\Theta_{b,p}$ from. This paper analyzes several simple methods of predicting $\Theta_{b,p}$, and proposes a new method using collaborative filtering to produce more accurate predictions for $\Theta_{b,p}$, particularly in cases of sparse data.

purpose. Batting average is defined similarly to OBP, although it ignores any plate appearances which resulted in walks. For this reason, most modern baseball statistical analysis uses OBP

2. Data Set

This paper utilizes a set of data for current professional baseball batters and pitchers in Major League Baseball, crawled from the publicly available data on ESPN.com on November 22, 2006 [ESPN06]. The data set includes 788 batters and 862 pitchers, with the results of 911,871 plate appearances coming from 167,759 unique batter-pitcher matchups. One challenge to this data set is that the complete batter-pitcher matrix is relatively sparse, there is only data for $\frac{167759}{788 \cdot 862} \approx 25\%$ of all possible pairs. In particular, there are are many individual players and batters with very limited careers, including several with only one matchup. Second, the typical number of plate appearances available for a given matchup is quite low, the average is $\frac{911871}{167759} \approx 5.44$ plate appearances, the median (ignoring empty matchups) is just 4.

For the purposes of this paper, only data from matchups between currently active batters and pitchers is considered. For some older players, this means throwing out a significant amount of data from their career OBP which came in appearances against now-retired pitchers. An extreme example is 48-year old Julio Franco, who in reality has 9,490 career plate appearances but only 2,197 against currently active pitchers. Franco’s OBP is .358 against the current pitchers, as opposed to .365 against all pitchers he has ever faced (this difference is explained because most of the prime of his career was against now-retired pitchers). For this paper, we consider the first value to be Franco’s “career” OBP, ignoring the earlier part of his career.

3. The Unique Matchup Hypothesis

3.1 Description

Motivating the desire to build an accurate predictor of matchups is the underlying assumption that batters fare differently against different pitchers, that is, that $\Theta_{b,p}$ is not always be well-predicted by the values of Θ_b and Θ_p^{-1} . We will call this the *Unique Matchup Hypothesis*, more formally that, if $\Theta_{b,p}$ were known precisely for a batter b and a pitcher p , it would not be equal to a simple weighted average $\Omega_{b,p} = \omega \cdot \Theta_b + (1 - \omega) \cdot \Theta_p^{-1}$. Most baseball players and analysts would agree, intuitively, that this seems to be true. Different batters and pitchers have very different styles. Some pitchers, for example, throw mostly high-velocity fastballs, while others throw more curveballs or changeups. Different batters have unique skills in terms of hitting higher velocity pitches and hitting breaking balls or off-speed pitches. Every pitcher and every batter seem to be unique, and therefore it seems that any matchup should have unique results.

3.2 Handedness correction

One objection which may be raised at this point is that, although all hitters and pitchers are to some extent unique in their abilities, they both group into two major classes based on handedness, that is, batters and pitchers are either left or right-handed. It is a widely held baseball adage that left-handed hitters hit right-handed pitchers better than left-handed pitchers, and similarly that right-handed hitters hit left-handed pitchers better than right-handed pitchers. Virtually no baseball analyst would dispute this trend, but it is worth checking that it does clearly exist in the dataset at hand:

Table 1: Matchup statistics by handedness

Matchup	# plate appearances	Θ
LH batter, LH pitcher	64,373	0.320
LH batter, RH pitcher	321,690	0.348
RH batter, LH pitcher	178,771	0.344
RH batter, RH pitcher	347,037	0.321
Total	911,871	0.335

With the enormous number of plate appearances available, these differences are overwhelmingly significant from a statistical standpoint and reflect the traditional baseball wisdom well. Therefore, using a simple weighted prediction $\omega \cdot \Theta_b + (1 - \omega) \cdot \Theta_p^{-1}$ to estimate $\Theta_{b,p}$ is doomed to be inaccurate because it is ignoring the handedness of the matchup between b and p . To correct for this, we define a handedness-aware average

$$\Omega'_{b,p} = \omega \cdot \Theta'_b + (1 - \omega) \cdot \Theta_p'^{-1}$$

where Θ'_b is b ’s career OBP against all pitchers of the same handedness as p , and $\Theta_p'^{-1}$ is p ’s career OBPA against all batters of the same handedness as b . A minor complication is the existence of switch-hitters, who are able to bat ambidextrously. In all but extremely rare cases switch-hitters will choose to bat with the opposite hand as the pitcher they are facing, for the purposes of this paper switch hitters are considered right-handed when facing a left-handed pitcher and vice-versa, this is also how their statistics are figured into Table 1.

3.3 Statistical evidence

Statistically, it is difficult to prove beyond doubt that the unique matchup assumption is correct, especially when the handedness correction is used. The main problem is that most hitters and pitchers have faced each other an insignificant number of times (≤ 10) to draw any decisive conclusions. Also, the amount by which $\Theta_{b,p}$ differs from $\Omega'_{b,p}$ is

probably not large. As an example, most baseball followers would be comfortable with the claim that a certain batter b with $\Theta_b = .300$ can hit a certain pitcher p with $\Theta_p^{-1} = .300$ at a rate of $\Theta_{b,p} = .350$, if b happens to matchup particularly well with p . However, for any given value of $\Theta_{b,p}$ the standard deviation will be

$$\sigma = \sqrt{E(X^2) - E(X)^2} = \sqrt{\Theta_{b,p} - \Theta_{b,p}^2}$$

For $\Theta_{b,p} = .350$, this means the standard deviation is $\sigma \approx 0.477$, and thus the standard error for N trials will be $\frac{\sigma}{\sqrt{N}}$, meaning at least $N = 91$ trials are needed to reliably determine that $\Theta_{b,p} \neq \Omega'_{b,p}$. This is problematic because no professional batter and pitcher have faced one another 91 times, and most matchups differ from what would be guessed by less than 0.05.

Nevertheless, one can attempt to sort through the data and identify matchups for which $\Theta_{b,p}$ and $\Omega'_{b,p}$ differ by a significant degree. If we consider the null hypothesis to be that all matchups behave according to the distribution predicted by $\Omega'_{b,p}$, we can apply a G-test for statistical significance against the chi-square distribution. To do so, we first filter out all pairs b, p with fewer than 20 plate appearances, since the test is considered inaccurate for such low values. This limits us to 6,889 pairs of batters and pitchers. Next, we evaluate all pairs and can compute a statistical estimate of the likelihood that the difference between the observed value $\Theta_{b,p}$ and $\Omega'_{b,p}$ is not due to statistical noise. We can accept any pair as “significant” which exceeds some confidence threshold τ . Of course, by Bonferroni’s principle if we test N values we will expect to get $N \cdot (1 - \tau)$ incorrect significant values. Below are the expected and observed numbers of “significant” matchups at various values of τ :

Table 2: Occurrence of significantly unusual matchups

τ	expected # sig. matchups	observed #
0.9	688.9	775
0.99	68.89	108
0.999	6.889	15
0.9999	.6889	3

These results do support that the Unique Matchup Hypothesis is correct, in that more significant values are being observed than could be explained by statistical noise. In particular, the chances of having seen 108 values at 99% confidence are less than 10^{-5} , which is small enough to dismiss.

4 Evaluation of Simple Methods

4.1 Description of methods

It is worth discussing what simple methods are most likely used in practice to predict matchups. In reality, most professional baseball managers probably use no systematic method at all to decide between batters b_1 and b_2 in a game situation, but use a combination of gut feelings and some statistical information. We will assume a rational manager will make all decisions by defining some approximation for $\Theta_{b_1,p}$ and $\Theta_{b_2,p}$ and choosing the larger value. In this case, the problem is reduced to coming up with an accurate prediction method for $\Omega_{b,p}$ given some b and p . The simplest method which we can use as a baseline is always predicting the league-wide average OBP, $\bar{\Theta} \approx 0.335$. This method of constant prediction provides no decision making power but can be analyzed in terms of its accuracy compared to reasonable methods.

The next method, as discussed previously, is a simple weighted average $\Omega_{b,p} = \omega \cdot \Theta_b + (1 - \omega) \cdot \Theta_p^{-1}$. Many managers probably do this subconsciously, and for decision making between b_1 and b_2 it reduces to simply comparing Θ_{b_1} and Θ_{b_2} . As introduced above, a significant improvement is possible by instead using the handedness-corrected $\Omega'_{b,p} = \omega \cdot \Theta'_b + (1 - \omega) \cdot \Theta_p^{-1}$.

Finally, it is possible to incorporate some past experience between b and p , if available, to predict the matchup. Suppose the value $\Theta_{b,p}$ has been observed in k previous matchups. How much should this information be weighed into a prediction? Clearly, for low k , $\Theta_{b,p}$ is relatively meaningless and $\Omega'_{b,p}$ should still be used. Some prediction $\Lambda_{b,p}$ combining $\tilde{\Theta}_{b,p}$ and $\Omega'_{b,p}$ should be made. One simple possibility is an exponentially decreasing weight based on k , such as

$$\Lambda_{b,p} = (1 - \beta^k) \cdot \tilde{\Theta}_{b,p} + \beta^k \cdot \Omega'_{b,p}$$

This seems to work reasonably well with $\beta \approx 0.9999$. Another approach is a simple weighted average:

$$\Lambda_{b,p} = \frac{k}{k + \alpha} \cdot \tilde{\Theta}_{b,p} + \frac{\alpha}{k + \alpha} \cdot \Omega'_{b,p}$$

Empirically this works slightly better, and has a nice real-world explanation in that it is equivalent to adding α plate appearances which behaved according to the weighted average given by $\Omega'_{b,p}$. The best value of α is surprisingly high, around 300. This is extremely conservative in accepting information from $\tilde{\Theta}_{b,p}$ in the face of low k .

4.2 Evaluating the methods

To evaluate these methods, and the collaborative filtering techniques as they are introduced, they can be used to pre-

dict the outcome of every plate appearance in the data set by removing each value, predicting the result, and then computing the Root Mean Squared error. This is the most rigorous way to test the prediction schemes given the data available, and it especially penalizes values which are incorrect by a large amount. The RMS values for the simple schemes above are given below:

Table 3: Comparison of simple methods

Prediction Scheme	Symbol	RMS error
Constant prediction	Θ	0.4722006
Weighted Average	$\Omega_{b,p}$	0.4706462
Hand-corrected Average	$\Omega'_{b,p}$	0.4700132
Historical Extrapolation	$\Lambda_{b,p}$	0.4699862
“Perfect” Prediction	–	0.4246710

The problem with RMS is that the values have no inherent meaning except as a relative measure of success. “Perfect” accuracy is added which represents an engine which can guess the optimum values by knowing the test set in advance. This method is cheating because it uses the unknown data in prediction. It only serves to provide a lower bound on the error given by any deterministic prediction scheme, based on the variance of the data. Still, it can be seen that constant prediction is poor, slicing data by handedness is helpful, and using historical data is surprisingly unhelpful. The reason for this is that because most matchups have a meaningless amount of historical data, it has to be weighted very low or else it will cause strong errors for matchups with a low number of previous occurrences k .

5 Collaborative Filtering

5.1 Background

Collaborative filtering is a technique which is primarily used by online retailers to recommend products and services to customers. The classic example is a movie recommendation system. Since individuals have different tastes in movies, a one-size-fits-all recommendation system is not sufficient. However, because there are natural clusters of users who like similar movies, such as those in a particular genre or with a particular director, individualized recommendations can be given to a user u by finding a set S_u of users who rate movies similarly to u , and recommending all the movies highly rated by members of S_u to u .

Although baseball prediction seems to be a very different problem from movie recommendations, in fact the same technique can be applied. Generally, to predict how well b will do against p , we want to find a set of similar batters to

b , and use their collective performance against p to predict b ’s performance. This approach is appealing because it has the potential to solve the problem of insignificant history between b and p . By identifying a large enough set S_b of sufficiently similar batters, hopefully there will be enough (ie, several hundred) plate appearances between all members of S_b and p that a more meaningful conclusion can be drawn.

5.2 Approach

We want to consider a batter b in terms of his matchup vector $\vec{b} = [\Theta_{b,p_1}, \Theta_{b,p_2}, \dots, \Theta_{b,p_n}]$ for all n pitchers in the data set. There are a few issues to address with this model. First of all, there is probably no information between b and most pitchers. As discussed in Section 2, over 75% of all matchups have no data. This is not necessarily a problem, though, as movie-recommendation systems have been shown to perform adequately despite fewer than 1% of values being known. A bigger problem is that, even where data exists, we don’t know Θ_{b,p_i} precisely but only an approximation based on a small set of past results. In the movie-recommendation scenario, all known values can be relied on to be true, if a user has given a movie three stars then that truly reflects the user’s opinion of the movie.

In the case of baseball, however, simply using the observed value $\tilde{\Theta}_{b,p}$ in the place of each $\Theta_{b,p}$ will result in a vector for b with many extreme values, such as against pitchers b has only faced once or twice. To correct for this, we must attenuate each component of the vector based on the number of plate appearances used to generate it. Similar to the formula for making a weighted historical decision, we replace $\tilde{\Theta}_{b,p}$ with:

$$\Lambda_{b,p} = \frac{k}{k + \alpha} \cdot \tilde{\Theta}_{b,p} + \frac{\alpha}{k + \alpha} \cdot \Omega'_{b,p}$$

However, it is important to use a much lower value of α in this case, but we do need to preserve significant variation between different hitters. Experimentally, $\alpha = 20$ seems to work fairly well. This has the overall effect of adding to each component of the vector 20 plate appearances at which b performed at exactly the rate given by $\Omega'_{b,p}$.

5.3 Computing Similarity

Now that we have constructed a vector \vec{b} describing b ’s performance against the various pitchers, we must define a way to calculate similarity between two batter’s vectors \vec{b}_1 and \vec{b}_2 . Two methods are commonly used in collaborative filtering applications, cosine similarity given by a normalized vector dot-product and Pearson’s correlation coefficient. In

our case, we will use the former, and define

$$\text{sim}(\vec{b}_1, \vec{b}_2) = \frac{\vec{b}_1 \cdot \vec{b}_2}{|\vec{b}_1| |\vec{b}_2|}$$

A further restriction imposed is that no two batters are ever considered similar if they have not faced some minimum number of common pitchers, such as 10. This avoids assigning high similarity to two batters without adequate support in the data. It also has the effect of removing the ability to make any predictions about some batters who have faced less than 10 total pitchers in their career, although this is a relatively small number.

5.4 Normalizing the vectors

We may choose to normalize the vectors \vec{b} before using them to compute similarity. For example, we might want to divide each component Θ_{b,p_i} of \vec{b} by the prediction given by Ω'_{b,p_i} . This will change \vec{b} from describing the absolute success of b against various pitchers into describing how much better or worse b does than could be expected based on b and p 's abilities. The batters identified as being similar to b are not likely to be of the same relative skill as b . Many of the similarities produced if the vector is normalized in this manner make little sense on the surface from a baseball perspective. For example, Alex Rodriguez, considered one of the dominant hitters of the current generation, is considered most similar to Josh Rabe, an obscure left-fielder with 51 career plate appearances. Surprisingly, though, this form of normalization makes the predictions significantly more accurate. It speaks to the general power of the method that it is able to successfully utilize similarities which a human baseball analyst would otherwise never consider. Normalizing with Ω'_{b,p_i} appears to be the best method, although the non-handedness adjusted Ω_{b,p_i} was tested as well.

5.5 Making predictions

Given the above machinery, we are able to produce for each batter b an ordered list $\{b_1, b_2, \dots, b_m\}$ of all other batters, ranked by their similarity to b . Now, to predict b 's success against p , we want to examine the results when some group of batters similar to b faced p . We will compute the average of their individual results $\Theta_{b_i,p}$ to produce $BB_{b,p}$, the batter-batter collaborative filtering estimate for b against p . It must be decided how many additional batters are to be considered. What has seemed to work the best is to determine a value μ of plate appearances desired, and add the results of the similar batters b_1, b_2 , etc. against p until a total of μ plate appearances have been considered. This guarantees that a reasonably large number of matchups are being considered. If instead a constant number N of batters are examined, the estimate might be poor if the pitcher is new

or rare, and the N most similar batters have a small number of appearances against p . The value $\mu = 280$ was found to be the most effective experimentally.

It is also necessary in this step, if the vectors have been normalized, to "un-normalize" them. For example, if normalization by $\Omega'_{b,p}$ is used, and we are adding the results of k appearances between some similar batter b_1 and p , we must multiply the value $\Theta_{b,p}$ by $\frac{\Omega'_{b,p}}{\Omega'_{b_1,p}}$. This corrects for the fact that we are adding data from batters who may vary in their overall ability from b .

5.6 Pitcher-Pitcher filtering

All of the above has focused on batter-batter filtering, that is, finding batters similar to b and examining their performance against p . The reverse process can also be done, namely, looking for pitchers similar to p and then examining their history against b . The details are almost entirely symmetrical to those described above for batter-batter filtering, including the use of normalization to improve results. For making predictions, the value $\mu = 320$ was found to be the best.

Pitcher-Pitcher filtering also proved more powerful than batter-batter filtering. From a baseball perspective, this implies that pitchers cluster better than batters do, that is, they are more easily divided into similar groups. This observation is reasonable from a baseball standpoint, in that pitchers do fall into a number of similar categories mainly based on the velocity with which they throw and what types of pitches they throw.

5.7 Combining methods

To build a strong prediction engine, results are combined and weighted from both batter-batter filtering, pitcher-pitcher filtering, and the simple statistical engines previously described. This strategy of combining results into a hybrid scheme is similar to that used in recommendation-systems, which combine results from similar users with "content-based" recommendations, which are recommendations based on genres, actors, etc similar to those highly rated by a user. In our case, the statistical estimate $\Omega'_{b,p}$ represents a content-based estimate, and we now also have the collaborative-filtering estimates $BB_{b,p}$ and $PP_{b,p}$. These can be combined in any weighted calculation

$$H_{b,p} = \gamma_1 \cdot \Omega'_{b,p} + \gamma_2 \cdot BB_{b,p} + \gamma_3 \cdot PP_{b,p}$$

Once again, the γ values must be calculated experimentally. The best results were seen on this data set with $\gamma_1 = 0.3$, $\gamma_2 = 0.25$, and $\gamma_3 = 0.45$.

In fact, this is the only way that the collaborative filtering estimates can be used, since for new batters and pitchers the

calculation will find no sufficiently similar players and be unable to make any predictions. This is similar to the “new user” problem experienced in recommendation engines. In this case, the hybrid scheme handles the situation gracefully by defaulting to the handedness-adjusted weighted estimate.

5.8 Cost

As described and implemented, computing the similarity of each of N batters with one another, as well as the cost of computing the similarity of M pitchers with one another, has a worst-case running time of $O(M \cdot N^2 + N \cdot M^2)$. This would not scale well to large M and N , and in other implementations of collaborative filtering methods there are more efficient techniques to approximate this calculation. However, since $M \approx N \approx 800$, this calculation is still fairly easy. Scalability is not a major concern because there simply do not exist larger data sets of interest for this type of calculation. Also, the larger calculation can be performed as pre-processing, and the real-time cost of answering individual queries about batters b and p is then $O(M + N)$.

6. Results

6.1 RMS Error

The results of the best hybrid collaborative filtering scheme against simpler methods are presented in Table 4. Results are shown both for the unnormalized method, and with normalization to Ω'_{b,p_i} .

Table 4: Comparison of methods

Prediction Scheme	Symbol	RMS error
Constant prediction	Θ	0.4722006
Weighted Average	$\Omega_{b,p}$	0.4706462
Hand-corrected Average	$\Omega'_{b,p}$	0.4700132
Historical Extrapolation	$\Lambda_{b,p}$	0.4699862
Coll. Filtering (unnormalized)	$H_{b,p}$	0.4668856
Coll. Filtering (normalized)	$H'_{b,p}$	0.4663412
“Perfect” Prediction	–	0.4246710

Again, it is difficult to exactly grasp the meaning of these results, because RMS is a relatively abstract concept. However, it is instantly clear that the benefit of a collaborative filtering approach is tremendous, in that the error is lowered from a constant prediction scheme by almost three times as much as using a simple handedness-adjusted scheme, or even historical extrapolation. The specific numbers are not important, especially because all of the above numbers could probably be improved slightly with tweaks to the various formulae. The important thing is that collaborative fil-

tering appears to do significantly better in terms of reducing the raw error.

6.2 Real-world impact

Because the RMS value is only useful to compare the relative performance of prediction schemes, it is worth asking the question of how much the improvements of collaborative filtering could realistically impact the success of a team in terms of wins and losses. Ultimately, the only utility of a matchup prediction scheme is in its ability to aid in decision making. To try and estimate the importance of predicting matchups, we can compare the decisions made by the best collaborative filtering estimate $H'_{b,p}$ with the best simple statistical estimate $\Omega'_{b,p}$. We will try to examine what the benefit will be of a manager making decisions using collaborative filtering.

To do so, we consider all pairs of batters b_1, b_2 which have both had at least 25 appearances against a common pitcher p . We can examine what decision is made by the two systems in picking between b_1 and b_2 , without using any of the observed data $\Theta_{b_1,p}$ or $\Theta_{b_2,p}$. In every situation where the two systems would lead to a different decision, we add the difference $\Theta_{b_H,p} - \Theta_{b_{\Omega},p}$ where $\Theta_{b_H,p}$ is the batter suggested by using $H'_{b,p}$ and $\Theta_{b_{\Omega},p}$ is the batter suggested by using $\Omega'_{b,p}$. After dividing by the total number of decisions considered, this should produce a rough estimate of the average benefit of using $H'_{b,p}$.

This simulation was run over 200,825 pairs of batters b_1 and b_2 . Of these decisions, $H'_{b,p}$ and $\Omega'_{b,p}$ selected different batters for 38,113 of them, or 18.9%. Of these differences, $H'_{b,p}$ selected the batter with the higher actual value of $\Theta_{b_2,p}$ in 78.6% of cases. Already, this data suggests some utility to using collaborative filtering, as it guessed correctly in the majority of disagreements and was correct in most of these.

Most importantly, the average improvement in the OBP of the selected hitter over all decisions was 0.0156 for $H'_{b,p}$. This value means that a manager using the collaborative filtering approach to predict matchups could expect his team to gain an extra 1.8 hits for every 100 decisions made.

How much could this affect a baseball team’s fortunes? The answer depends on how many “decisions” are made by the manager in the course of the year, which is impossible to quantify. A very rough estimate is that a manager makes 5-8 decisions per game, including decisions about pinch hitters, decisions about relief pitchers, and decisions about the starting lineup². Over a 162 game season, then, it is realistic to expect a manager to make 1,000 decisions. Thus, a very rough estimate is that the collaborative filtering method, if applied, might result in 15 additional hits over the course of an entire season. Once again, it is difficult to conclusively

²Starting lineup decisions are especially important, because the selected hitter usually gets three or more plate appearances

say what this would amount to, but it is certainly realistic to think that this could mean 1 or 2 additional games won in the course of the year. Given that baseball is a major business in which each win victory costs on the order of \$1,000,000 in terms of player salary alone, it is certainly reasonable to think that the approach is worth adopting.

7 Possible Improvements

7.1 Consideration of more outcomes

One major limitation of this study is only considering on-base percentage. In reality, managers also care about the rate at which other outcomes occur, in particular home-runs and other extra base hits. All of these rates could theoretically be predicted in a very similar fashion. The only major issue is lack of data, since increasing the number of outcomes observed greatly reduces the data for each outcome. An alternative may be to try to predict a single representative statistic which better accounts for extra-base hits, such as OPS or Equivalent Average.

7.2 Consideration of time

This study has also completely removed the temporal dimension from the data, assuming a model where all batters and pitchers are identical throughout their career. This is of course, not accurate, in reality player's effectiveness changes significantly over time, and this must be taken into account when making real baseball decisions. Another improvement would be to consider a matchup prediction system which more heavily weights recent data than data which is further in the past.

8 Conclusion

Collaborative filtering is a powerful technique which applies well to predicting baseball matchups. The performance of collaborative filtering in an experimental setting appears to decisively beat simpler methods, to the point that it could actually influence a real team's win-loss record if used to make all matchup decisions. Because the cost of calculating it is relatively low, and there is a fairly high amount of monetary value to success in professional baseball, it seems that collaborative filtering should be widely adopted after further refinements.

Whether or not this will actually happen is an open question, given baseball's track record of resistance to modern statistical analysis. Despite the documented success of some modern teams in using statistical methods to better evaluate player talent, many organizations are reluctant to go against traditional baseball thinking. A prime example is

that many highly-paid baseball analysts and managers still do not accept the superiority of statistics like OBP and OPS to the more traditional batting average, despite overwhelming evidence that they are better predictors of baseball success. Still, collaborative filtering is clearly a worthwhile pursuit for baseball decision makers. If packaged properly into a simply designed tool, it should be a useful aid for managers in making in-game tactical decisions.

References

- [AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In *IEEE Transactions on Knowledge and Data Engineering*, pages 734-49, 2005.
- [LSY03] Greg Linden, Brent Smith, and Jeremy York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. In *IEEE Internet Computing*, pages 76-80, 2003.
- [ESPN06] ESPN.com Major League Baseball statistics <http://www.espn.com/mlb> November 22, 2006.