Data Driven Pitch Selection: Applying Machine Learning to Pitch Strategy Project Progress Report Version

Joey Asperger[1], **Jorge Garcia Mesa**[1]

[1]Stanford University
450 Serra Mall, Stanford, CA 94305

{joey2017,jgarci10}@stanford.edu

Abstract. *Given the the availability of data for Major League Baseball (MLB) games, baseball has become a hotbed for statistical analysis. In this paper we aim to develop better pitcher strategy within a game given statistical information about the batter. By applying common machine learning regression techniques to pitchFx data, we hoped to learn about sequence outcomes given features about the pitch and batter* [2, 3].

1. Introduction

Baseball, a bat-and-ball game is recognized as America's past time. Although, it holds this nickname its reach includes Latin America, the Caribbean, East Asia, and other parts of the globe. Within the game one of the key elements determining success is the dual between pitcher and batter. The pitcher throws that ball and the batter tries to hit it. A pretty simple concept, but a lot of strategy is involved in how the pitcher will throw and how the batter reacts. In this paper, we will use pitch analysis to develop a pitching strategy given different types of batters.

Using a machine learning algorithm, stochastic gradient descent, we incorporated information about the features of a pitch as well characteristics of the batter being faced. During testing, given batting statistics for batters, we classified the players into different categories. Given the classification, then we can provide useful information to pitcher on how to combat the batter.

2. Data

The dataset used came from the pitchFx. This is a system first installed in 2002 that is now present in all MLB stadiums. The system keeps track of the speed and trajectory of baseball pitches. Features of the pitches being tracked include speed of the ball, location of the pitch, break angle, vertical displacement, horizontal displacement, among others. Along with characteristics of the pitch, we are also provided IDs of the batter being faced. The batter IDs on pitchFx are then paired up with season batting statistics, which we used to characterise the batters [6]. Management of the database was done with the help of MongoDB.

3. Model

3.1. Challenges

Baseball like any other sport is very complicated when examined for what produces successful outcomes. In many ways, sports and particularly those with quantitative data are

gold mines for data analysis and machine learning. The problem comes in the fact that given so much data to work with, it can be difficult in choosing what is important to best simulate the real world without over-complicating the process.

3.2. Feature Selection

The pitchFX data has a lot of features that we could have worked with, but we narrowed down on what we thought were useful data points for the types of pitches we examined: four seam fastball, curveball, and changeup.

3.2.1. Four Seam Fastball

The four seam fastball is a part of the fastball family and is generally the fastest pitch a pitcher throws. For reference, at the major league level this pitch can reach speeds up to a 100 miles per hour. The pitch is gripped with the index and middle fingers being perpendicular to the seams [5].Designed for velocity with little break, this was emphasized in our features.

Features for Fastballs
start speed
horizontal location
vertical location
vertical displacement
horizontal displacement

Given these recorded features of a pitch we worked to extract more impact features. To do this we looked at the values relative to each other or a target, absolute values instead of relative, and features in relation to handiness [4].

3.2.2. Curveball

With a curveball, a pitcher will place the middle finger parallel to one of the long seams. The thumb will be just behind the seam on the opposite side of the ball. The grip and hand movement provides it with its characteristic downward motion when approaching the batter [5].

Features for Curveballs
start speed
horizontal location
vertical location
vertical displacement
horizontal displacement
break angle
break length
vertical break distance

Along with the features and variations that we had for fastball, in order to account for the desired change in trajectory (break), we also incorporated break features for the curveball and their variations.

3.2.3. Changeup

A type of off speed pitch. The pitch is usually intended to look like a fastball, but travel at a much slower speed. The changup is usually gripped with three fingers instead of two like a fastball and is also held farther back within the hand. This will make the throw slower and thereby hopefully confuse the batter's timing [5].

Features for Curveballs
start speed
horizontal location
vertical location
vertical displacement
horizontal displacement
break angle
break length
vertical break distance

For changeups we felt that although fairly different from the curveball, it should still include the break features and their variations given the fact that the slower speeds will result in more movement than a fastball.

4. Algorithms

4.1. Characterization of Batters

Before conducting classification on the data, we wanted to classify the type of batters that the pitcher was against. To do this we scavenged for thoughts on batter classification as well as using our own sports intuition. We concluded that the four classifications would be high volume power hitter, balanced power hitter, defensive high average, and in play high average. Their characteristics are in the next graph.

Type of batter	Characteristic 1	Characteristic 2	Characteristic 3
high volume hitter	20 home runs $\leq$	0.270 batting average $\geq$	none
balanced power hitter	20 home runs $\leq$	0.270 batting average $\leq$	none
defensive high average	0.270 batting avg $\leq$	walk percent ≥ 10	none
in play high average	0.270 batting avg $\leq$	walk percent ≥ 10	strikeout percentage ≥ 16

In the testing situation, it will necessary to approximate the batter to one of the groups. To do this, we looked at the statics of batter that the pitcher was facing during testing of the model. We did the characterization by giving weights to stats that we initially used to characterize the learning data (the processor from above). The relative weights allowed us to best aproximate the type of batter being faced.

4.2. Regression algorithm and implementation

The machine learning algorithm that we choose to use was stochastic gradient descent. In stochastic gradient instead of looping over all the examples before making an update, we would be making making an update after very new seen data point. To work with the

optimized algorithm for gradient descent, we used the scikit python package. We thought that to get the most out of the project, the focus should be on how we manage the data.

In the implementation of the algorithm the feature vectors already described before were used and the results were the outcomes of the batting sequences. These outcomes were given numerical values representing whether they are perceived as positive or negative outcomes. These assigned values then allow us to compare how well our regression algorithm was from the actual results given a testing data-set. The reason for wanting to test this is that if were going to provide advice on how to face specific types of hitters, we would want our predictions to be close to the actual outcomes.

5. Results

After running our regression model using the pitch data against different types of hits and with different types of pitches we were able to obtain the weights:

Four Seam Fastballs:feature weights relative type of batter

Feature	Power	Balanced	Defensive	In-play hitter
start speed	0.0134	0.0204	0.0211	0.0142
horizontal placement	0.0227	0.0456	0.0446	0.0680
horizontal placement (squared)	0.0498	0.0386	0.0494	0.0164
vertical placement	-0.0394	-0.0163	-0.0152	-0.0442
vertical placement (squared)	0.0577	0.0133	0.0733	0.0778
horizontal movement	-0.0022	0.0016	-0.0152	-0.0055
horizontal movement (squared)	0.0159	0.0013	0.0198	0.0085
vertical movement	0.0087	0.0444	-0.0306	0.0150
vertical movement (squared)	0.0419	0.0053	0.0626	0.0170

Curveballs: feature weights relative batter types

Feature	Power	Balanced	Defensive	In-play hitter
start speed	0.0086	-0.0179	0.0140	-0.0099
horizontal placement	-0.0194	0.0151	-0.0031	-0.0207
horizontal placement (squared)	0.0446	0.0333	0.0699	0.0464
vertical placement	-0.0804	-0.0672	-0.0490	-0.0939
vertical placement (squared)	0.0601	0.0301	0.0551	0.1140
horizontal movement	0.0020	-0.0458	-0.0216	-0.0002
horizontal movement (squared)	0.0157	-0.0382	0.0074	-0.0019
vertical placement	0.0508	0.0432	0.0032	0.0198
vertical placement (squared)	0.0537	0.0492	0.0132	0.0271
break angle	0.0095	-0.0004	-0.0349	-0.0241
break length	0.0038	0.0178	-0.0075	0.0136
break y	-0.0236	-0.0541	-0.0362	-0.0036

6. Discussion

Now that we have weights, we can begin to have a better understanding of how we might be able to create better pitching strategies for pitchers. The reason for this being that

weights give indication as to how to best improve the outcome. The weights provide us with magnitudes as well as direction.

The weights for the different types of pitches as well as the expected outcomes with different types of pitches will be our groundwork for advising on pitch selection. Therefore given knowledge about the hitter, the batter can be grouped into his closest batter category. Provided a batter category, the pitcher would have knowledge to what types of pitches are effective against that specific type of hitter as well as what characteristics to emphasize when throwing the pitch.

7. Conclusion

Out model still has a ways to go before achieving the level of work we hoped for. We have to continue working on hanging large portions of the pitching and batting data. An example of this being that categorization for a batter only occurred once based off one season. There are clear challenges when working large data-sets in sports. One has to dig through the data to find what is important while being careful of not oversimplifying the event.

8. Future Work

For right now the way in which the batters were classified was through a manual placement of their characteristics into what we perceived as distinct type of players: high volume power hitter, balance power hitter, defensive high average, and in play high average. During testing the batter faced would be classified as the closest approximation. For a variation during the next iteration of the project, we were hoping to look into K-Means formed groups of batters and how that might help.

For management and easier to analysis reasons we did not incorporate all individual seasons to classify batters at this time. We made the assumption that the batters would be relatively stable, but on the next iteration we will take into account that the batter might change.

Given weights for different types of pitches against different types of pitches, how might we be able create pitch sequences that would aid the pitcher? A thing to consider would be how we would manage how sequences of pitch types and locations effect the success of a pitcher-batter interaction.

Referências

[1] Gergory Donaker. *Applying Machine Learning to MLB Prediction & Analysis*. 2005.

[2] Gartheeban Ganeshapillai, John Guttag *A Data-driven Method for In-game Decision Making in MLB*. MIT Sloan Sports Analytics Conference 2014.

[3] Gartheeban Ganeshapillai, John Guttag *Predicting the Next Pitch*. MIT Sloan Sports Analytics Conference 2012.

[4] Isaac Laughlin *The Data Science Behind Baseball Pitching Strategy*. Galavine 2015.

[5] Josh Walsh *Pitch Identification Tutorial*. The Hardball Times 2007.

[6] Brooks Baseball Web. 15 Oct. 2015. $\langle http://www.brooksbaseball.net/pfxVB/pfx.php \rangle$