Using Machine Learning Algorithms to Identify Undervalued Baseball Players

Tatsuya Ishii (twishii)

I. Introduction:

Across 30 teams in Major League Baseball, the prize goal of data driven front offices is to identify key players that provide value to teams. There are two main ways to achieve this: through draft and development of amateur players and through free agent acquisition. The former is a much more difficult problem, with insufficient data from amateur games and vagaries of player development. The latter is comparatively simpler, but acquiring widely-known quality players through free agency generally devolves into bidding wars, with inflated contract figures that diminishes the per dollar contribution to the team. And for well-known players, the cost in terms of prospects via trade is also prohibitively high. However, one of the most cost effective ways of improving team performance is to identify players that are being undervalued by the game.

The Houston Astros is among the most data oriented front offices in baseball and they have been employing a data driven evaluation of players with noted success [6]. Most famously, when Collin McHugh was let go by his previous team, it was the Astros that signed him because they noticed that the spin rate on his curveball (2000 RPM) was much higher than the average curveball (1500 RPM) [9]. What may seem like an eccentric reason to favor a player from other teams' perspective paid off as McHugh went on to have a successful season with his new team. This year, the Astros signed Charlie Morton, who is also noted to have an above average spin rate on his curveball [10]. In similar vein the Philadelphia Phillies in a move for upside quickly signed Andrew Baily who, despite the checkered injury history in recent past, had the top spin rate on his fastball [10].

In the realm of public discussion on advanced baseball analytics, people are increasingly applying machine learning techniques to baseball data. Vince Genaro presented his findings on hitter performance against pitchers grouped by clusters in SABR Analytics Conference in 2013 [4] [5]. CMU Tartan Sports Analytics featured an application of cluster analysis on pitchers by handedness [11]. And Fangraphs Community Research recently published a piece presenting the use of clustering algorithms to group pitchers with similar repertoires [2]. Each of these works highlights promising results as well as directions for future research and works to advance the public knowledge on baseball analytics.

To continue expanding the frontier of data driven baseball, we explore machine learning approaches to identifying undervalued players. The central hypothesis for these undervalued players is that they are players whose process is good, but are not getting the results. The goal is to utilize clustering algorithms to identify players who are similar to successful players in terms of process. In practical terms, the objective is to find players who have struggled but hold the most potential relative to their process and are prime targets for acquisition. Finally, in the goal of obtaining an end to end model, we incorporate information obtained in the clustering analysis to build a predictive model to forecast player performance in the subsequent season.

II. Data:

The data for this project come from BaseballSavant, which hosts data from both PITCHf/x and Statcast [1]. As the wealth of information from BaseballSavant is richer for pitchers than it is for hitters, the project naturally focuses its attention on evaluating pitchers in a process-centric outlook. Some of the key measures we looked at include pitch movement, break angle, spin rate and release extension. We aggregate these features by pitch type for each pitcher. In total, 15 variables that capture pitcher process from BaseballSavant were examined. These features have been standardized to have mean zero and unit variance to allow for comparison on an equal basis.

For this project we excluded the outcome of each pitch as it is hugely reliant on team defense. Once the ball is in play, the outcome is largely out of control of the pitcher. It would be inaccurate to attribute the success the pitcher enjoys from superior defensive teammates as his underlying abilities when comparing with another pitcher whose teammates may be on the field solely for his offense. So we proceed with our focus on pitch properties and factors that the pitcher can control.

To safeguard against any excessive bad or good lucks that the players may have encountered, the cutoff for inclusion in this study is 750 pitches. To put in context, assuming that a pitcher throws about 15 to 30 pitches in an inning, this translates to pitchers who have thrown at least 25 to 50 innings. The latter is also the MLB criterion for when rookie eligibility expires. That leaves 320 pitchers with which to evaluate their underlying performances.

Because Statcast was just recently rolled out in 2015, there is only full season data for 2015 and 2016. Since the results from clustering algorithms need data on 2016 performance results to evaluate its effectiveness, the scope of analysis for this project will be based on pitcher profile from the 2015 season.

IV. Method:

As the types of player acquisition moves made by the Astros and Phillies demonstrate, front offices are increasingly evaluating players on idiosyncratic features of a pitcher's repertoire. In the same spirit, we started our analysis by implementing K-means clustering, which seeks to minimize cluster centroid by assigning observations to its closest centroid and re-computing cluster centroid to be mean of its assigned points until convergence [8]. We implemented the algorithm for each individual pitch types and looked for undervalued players who are clustered with more established pitchers.

To pick the value for k, we first looked at the Gap statistic [7]. However, this measure failed to yield sufficiently large k that would lead to interesting and diverse clusters. The elbow method was similarly unfruitful [8]. At this point we switched to a heuristic of employing k = sqrt(n/2) as a starting point and using it as a guide for picking the right k based on context. For example, fastballs may have less diversity than curveballs, so the choice of k for respective pitches should be appropriately adjusted.

In evaluating the results of the K-means clustering, we computed the mean cluster Earned Run Average (ERA) and subtracted it away from the individual pitcher ERA's. Then sort by this ERA differential and look at pitchers who had the highest difference. The reasoning is that these are the pitchers who vastly underperformed relative to their cluster mean and should be prime candidates for rebound next season. And because of the struggles these players faced, they may be undervalued by other teams.

As a measure of pitcher improvements in the subsequent season, we looked at Fangraphs Wins Above Replacement (WAR) over Innings Pitched (IP) in 2016 [3]. WAR is a measure of how valuable a player is to his team. However, starters and relievers are used differently and accumulate WAR at different rates. And moreover, roles change from year to year based on team needs and circumstances. To gauge a player's value on equal footing, we divided WAR by IP for a measure of contribution on an innings basis.

To compare results with another clustering algorithm, we also implemented Hierarchical clustering with complete linkage and Euclidean distance as the dissimilarity measure. Also known as the furthest-neighbor technique, complete linkage agglomerative clustering computes pairwise dissimilarity based on maximal intercluster dissimilarity and fuses pair of clusters that are the least dissimilar [7]. This specification gave the best result among other linkage choices (average, single, centroid) and dissimilarity measure (correlation).

In determining which specification was superior to another, we looked at the quality of the top ten players identified to be undervalued and the relative ranking of these players with another. For each option, the comparison metric was the sum of each identified player's WAR/IP divided by its rank in the top ten and chose the specification that had better overall results. In practical terms, the ranking provided by the algorithms could be used as players to target in trade talks with other teams and algorithms that reliably yield better quality players with more potential at the top would be highly valuable to organizations.

IV. Discussion:

Looking at the top ten pitchers that come up according to the highest ERA differential in each pitch type analysis, the K-means algorithm successfully identified players that indeed went on to have better seasons in 2016. From the perspective of the front office, these are the players whose upside should be gambled upon. Table 1 summarizes the top performing pitchers identified in each analysis by pitch type. On a per innings basis, each of these players rebounded nicely in 2016 from their down years in 2015. And overall, the algorithm identified intriguing players with some noted for their potential, but with inconsistent results.

Table 1: K-means Clustering Top Performing Players by Pitch Type

Pitch Type	Player Name	Differential	WAR/IP 2015	WAR/IP 2016	WAR_2016
FF	Mike Foltynewicz	2.31	-0.001	0.011	1.3
SL	Shane Greene	2.94	0.001	0.020	1.2
CU	Shane Greene	3.20	0.001	0.020	1.2
CH	Shane Greene	2.68	0.001	0.020	1.2
FT	Matt Garza	1.75	0.004	0.014	1.4
SI	Chris Rusin	1.24	0.007	0.019	1.6
FC	Chris Tillman	1.65	0.010	0.014	2.4
FS	Jeff Samardzija	1.18	0.012	0.013	2.6
KC	Chris Tillman	1.16	0.010	0.014	2.4

The next step of the analysis was more ambitious: to identify undervalued pitchers according to some specified repertoire mix. Specifically we looked for undervalued pitchers with the standard four pitch mix of a starter and various two pitch combinations. The results are in Table 2.

Table 2: K-means Clustering Top Performing Players by Repertoire

Pitch Type	Player Name	Differential	WAR/IP 2015	WAR/IP 2016	WAR_2016
FF/CH/CU/SL	Shane Greene	2.58	0.001	0.020	1.2
SI/SL	Chris Rusin	1.14	0.007	0.019	1.6
FT/FC	Matt Moore	1.90	0.003	0.011	2.2
SL/FC	Matt Moore	1.96	0.003	0.011	2.2

Again, the K-means algorithm was able to identify intriguing pitchers who struggled in 2015 but rebounded in 2016. However with the exception of Matt Moore, all other pitchers identified in Table 2 had been found in Table 1 with a simpler approach. While this is an important finding to calibrate our approach, it is difficult to draw proper baseball conclusions as each pitch plays off one another.

Next, we examine players identified from the Hierarchical clustering algorithm. The results are listed in Table 3. The top performers from both algorithms in each pitch category show a lot of overlap and this is also the case with the repertoire approach.

Table 3: Hierarchical Clustering Top Performing Players

Pitch Type	Player Name	Differential	WAR/IP 2015	WAR/IP 2016	WAR_2016
FF	Shane Greene	2.96	0.001	0.020	1.2
SL	Shane Greene	2.98	0.001	0.020	1.2
CU	Chris Rusin	1.96	0.007	0.019	1.6
CH	Matt Moore	1.98	0.003	0.011	2.2
FT	Matt Moore	1.77	0.003	0.011	2.2
SI	Chris Rusin	1.42	0.007	0.019	1.6
FC	Matt Moore	1.87	0.003	0.011	2.2
FS	Jeff Samardzija	1.19	0.012	0.013	2.6
KC	Chris Tillman	1.44	0.010	0.014	2.4
FF/CH/CU/SL	Matt Garza	1.73	0.004	0.014	1.4
SI/SL	Chris Rusin	1.14	0.007	0.019	1.6
FT/FC	Jeff Samardzija	1.1	0.012	0.013	2.6
SL/FC	Matt Moore	1.96	0.003	0.011	2.2

On one hand, it is reassuring that both algorithms identified similar set of intriguing players. But the next criterion would be to judge how effective the algorithms were at identifying these players. We focus on the top ten players identified to likely rebound next season by each of the algorithms and see how valuable these players were in the subsequent season and their relative ranking within the top ten. The measure used here to compare the two algorithms is similar to the one described in selecting the specification for the Hierarchical clustering. The results are shown in Table 4.

Table 4: Algorithm Player Identification Ranking Comparison

WAR/IP Delta			WAR/IP 2016			WAR 2016		
Pitch	KM	HC	Pitch	KM	HC	Pitch	KM	HC
FF	0.024	0.028	FF	0.009	0.011	FF	1.58	1.48
SL	0.035	0.037	SL	0.014	0.014	SL	1.10	1.24
CU	0.012	0.011	CU	0.006	0.003	CU	1.51	1.60
CH	0.024	0.034	CH	0.007	0.013	CH	1.00	1.50
SI	0.016	0.016	SI	-0.005	0.001	SI	0.38	0.58
FC	-0.018	-0.024	FC	-0.011	-0.016	FC	0.85	0.77
FS	-0.022	-0.022	FS	0.002	0.002	FS	1.27	1.27
KC	0.000	0.007	KC	0.020	0.015	KC	2.72	2.70
FT	0.002	0.006	FT	0.006	0.010	FT	1.50	1.86

In two of the three categories, Hierarchical clustering had better results in the top ten than K-means clustering. However, part of our conclusion depends on how we define undervalued to be. If the goal is to identify who was the most valuable on a per innings basis, then the K-means algorithm produced equally compelling results. On the other hand, if the objective is to identify who improved the most or was the most valuable, then Hierarchical clustering had the upper hand.

V. Extension: Predicting Performance Improvement

Utilizing the data on cluster assignments and ERA differential obtained in the clustering analysis, we make a transition from the unsupervised learning clustering analysis to supervised learning predictive modeling problem. Specifically, our goal will be to predict 2016 WAR figures from information based on data from 2015. While WAR/IP was a useful metric for discovering undervalued players, predicting WAR figures themselves would be of greater utility to front offices as it gives them better indication of how the team is likely to fare. The goal now is to project how the players identified as undervalued perform in the upcoming season.

For this task, we selected Boosting and Random Forest machine learning algorithms for their noted success in Kaggle competitions. To complement our results from the clustering analysis, we also introduced additional data on player outcomes such as Strikeouts, Walks, and Home Rates as well as Flyball and Groundball tendencies and looked at over 30 variables in total.

We computed the ten-fold cross validated absolute mean errors with the Boosting and Random Forest models performing 0.847 and 0.857, respectively. Moreover, the addition of data obtained in the clustering analysis had only a marginal impact on predictive performance.

VI. Conclusion:

In this paper, we explored the effectiveness of machine learning clustering algorithms and pitcher process approach to finding undervalued baseball players. By focusing on players whose ERA was vastly higher than their cluster ERA, both the K-means clustering and Hierarchical clustering algorithms identified intriguing players who bounced back in 2016 from their down years in 2015. We experimented with clustering based on both individual pitch and repertoire approaches and found that the former was likely to be as effective as the latter.

As an extension exercise, we also tried projecting player value in the subsequent season and this task proved to be more challenging. The results from the clustering analysis had minimal impact on the predictive power of the models. While we incorporated more measures on player outcomes to augment the model, we likely need further investigation on identifying the types of data necessary to forecast figures for player value. Although our approach to identifying who is likely to improve was sufficient with data on player process, we need more data on player profile to get a closer estimate of how much that improvement likely is.

Nevertheless, our initial exploration demonstrated the potential of process based approach to player clustering analysis. Further areas for investigation include using different measures such as cluster mean Fielding Independent Pitching deviation to identify possibly undervalued players. In addition, since each pitch type complements one another, there is more room to innovate the repertoire based approach to the clustering analysis. And as we accumulate more data in the Statcast era, the increasing range of analyses that becomes available is bound to shed light on exciting new baseball knowledge waiting to be uncovered.

References:

[1] BaseballSavant: https://baseballsavant.mlb.com/statcast_search

[2] "Clustering Pitchers With PITCHf/x." Fangraphs. October 18, 2016. Accessed November 20, 2016. http://www.fangraphs.com/community/clustering-pitchers-with-pitchfx/.

[3] Fangraphs: http://www.fangraphs.com/

[4] Gennaro, Vince. "Clustering Pitchers by Similarity: Part 1." Diamond Dollars. April 22, 2013. Accessed November 20, 2016. http://vincegennaro.mlblogs.com/2013/04/22/clustering-pitchers-by-similarity-part-1/.

[5] Gennaro, Vince. "Clustering Pitchers by Similarity: Part 2." Diamond Dollars. June 3, 2013. Accessed November 20, 2016. http://vincegennaro.mlblogs.com/2013/06/03/clustering-pitchers-by-similarity-part-2/.

[6] Green, Joshua. "Extreme Moneyball: The Houston Astros Go All In on Data Analysis." Bloomberg. August 28, 2014. Accessed November 20, 2016. http://www.bloomberg.com/news/articles/2014-08-28/extreme-moneyball-houston-astros-jeff-luhnow-lets-data-reign.

[7] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. New York, NY: Springer, 2009.

[8] Jones, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. New York, NY: Springer, 2013.

[9] Sarris, Eno. "Finding the Next Collin McHugh With Spin Rates." Fangraphs. October 3, 2014. Accessed November 20, 2016. http://www.fangraphs.com/fantasy/finding-the-next-collin-mchugh-with-spin-rates/.

[10] Sarris, Eno. "Using Spin to Identify Two Underrated Free Agents." Fangraphs. November 16, 2016. Accessed November 20, 2016. http://www.fangraphs.com/blogs/using-spin-to-identify-two-underrated-free-agents/.

[11] Silverman, Steven. "A New Method for Clustering Pitchers" Tartan Sports Analytics. March 2, 2016. Accessed November 20, 2016. https://tartansportsanalytics.com/2016/03/02/a-new-method-for-clustering-pitchers/.