European Journal of Physical Education and Sport Science
ISSN: 2501 - 1235
ISSN-L: 2501 - 1235
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.2548931 Volume 5 | Issue 4 | 2019

ANALYSIS OF FIELDER STARTS AND BENCH ABILITY ON AMERICAN PROFESSIONAL BASEBALL PLAYERS

Wen-Sheng Chiu[1],
Kuo-Wei Lin[2i],
Yen-Chieh Wen[3]
[1,2]Physical Education Office,
National Central University,
Taiwan (R.O.C.)
[3]Physical Education Office,
National Taiwan University of Arts,
Taiwan (R.O.C.)

Abstract:
The development of athletes or players depends on two aspects: nature and nurture. The former is the talent and qualification of the players themselves, while the latter is the training that consumes human, material and financial resources. Take professional baseball players as an example. Matching the talents of players and referring to the relevant starting rules of the professional baseball league, when the up-and-coming players are first discovered, focused training are used on them. By doing so, the value of the players would be effectively enhanced and the players are helped to seek a better way out. This can form a virtuous circle: the pellets get quality players, and the players get better results. That is to say, strengthening the training for the shortcomings of the players with the potential of the starting players can avoid unnecessary training and huge training expenses behind them, and greatly reduce the risk of career, so that the players have higher security in their short career, and get a win-win-win situation. This study is aimed at the schedule information of the American Baseball League teams. Through feature selection of data mining, this study analyzes the main relationships and key differences between starting player and bench player of second baseman and shortstop in League of Nations teams. It is found that the on base percentage and speed of the infielders is an important ability indicator for the starting position; whereas, the second baseman emphasizes on the attack and the shortstop focuses on fielding. This feature is verified by comparing the opinions of experts and commentators.

Keywords: data mining, feature selection, sports forecast, major league

[i] Correspondence: email fjcu70@hotmail.com

1. Introduction

1.1 Research Motivation and Purpose

The US Major League Baseball (MLB) is the most advanced baseball league in North America. The league was formed in 1902 by the merger of the National League and the American League. Thanks to the well-organized and successful operation of the business model, and the excellent salary which has attracted talents, the league has become the highest hall for baseball players around the world.

Different defend positions require players with different abilities, including foot long and slugging percentage, and even batting ability. Each pellet must carefully consider the players' personal traits for arranging and scheduling. For example, the fielders near the two sides (first-base man, third-base man, left-fielder, and right - fielder) value the power of batting, while the middle fielder (second baseman, shortstop, and center-fielder) pays attention to the fielding and speed. These offensive and defensive balances are a must for scheduling line up.

All players are working hard to become a fixed starter. The pitcher wants to be the main starting pitcher, and the fielder is one of the starting nine players whom the team heavily relies on. The starting position not only proves the ability to affirm the player, but also has the effect of stabilizing the player's performance and strengthening his ability. For pitchers, the difference in the number of starting pitchers and bench players is less meaningful, because the starting pitcher must obey the prescribed number of days off, and the bench player might have to go to the field in response to the situation. In contrast, the role of starting fielder is quite different from the role of the bench fielder. Initially, the starting players are the main players. As the game goes on, these starting players are adjusted according to the situation, and the bench players might have chance to join the game. Players' replacement considerations are mostly based on the essential difference between the starting players and the bench players, as well as the actual needs of the game. Therefore, the game played is directly related to the ability of the player.

This study analyzes the competition results of the second baseman and shortstop of the US Major League Baseball. Feature selection is used as a tool to explore the differences between the requirements of the ability of the bench fielder and the starting fielder. This study also clarifies the teams' ability to determine the starting position. The result is finally provided to be the numerical reference for the training of players, which is expected to achieve better performance to meet the needs of the starting. This research result could not only provide players with a clear and practical development and training direction, reduce the risk of the players' career, but also reduce the risk of misplaced training for the pellets, resulting in well-targeted and more powerful players. Finally, by comparing the opinion of experts and commentators, a reliability assessment of whether the starting feature is accurately depicted could be applied.

1.2. Research Scope and Limitations

The research sample is taken from the 2015 US Major League Baseball season and the player, Chin-Lung Hu, is used as a case study. Therefore, the ability and prediction model constructed is only applicable to the position of the US Major League Baseball as a shortstop or second baseman. The same model does not apply to all defend positions. Because of the different capabilities required for different garrison locations, it must be constructed according to the appropriate defensive position of the predicted object to augment the applicable object.

The data is taken from May 8th, and the subsequent data is used as the identification of accuracy. Since the topic of discussion is the relationship between two factors: the starting player and his ability, this study sets the two factors as independent variables and dependent variables. Described as follows:

a) Independent variables refer to the capabilities that must be possessed by the fielder, for example: fielding percentage (FPCT), runs batted in (RBI), hits (H), second baseman (2B), third baseman (3B), home runs (HR), runs (R), total bases (TB), bases on balls (BB), strike outs (SO), stolen bases (SB), caught stealing (CS), on base percentage (OBP), slugging percentage (SLG), batting average (AVG).
b) The dependent variables are the starting percentage, bench percentage and other attack and defensive performance. The starting percentage is defined as games started (GS) of the starting player divided by the number of times the team played. If the starting percentage is greater than 0.5, it is regarded as starting; if the team does not have any players' starting percentage greater than 0.5, then the one with the highest percentage is the starting.

According to the intensity of the game, the statistical records of the US Major League Baseball would be calculated separately from the results of the season and playoffs. Whether it is included in the playoffs as the basis for analysis depends on whether the number of documents reaches more than 30, and the analysis results are objectively representative. In addition, the starting of the pitcher, the backup and the fielder is fundamentally different, and the independent variables of abilities are distinct. Therefore, this study is not suitable for the analysis of pitchers.

2. Literature Review

2.1. Baseball Related Prediction Literature

Baseball-related prediction studies in the literature could be divided into two categories. The first category is to predict or analyze the win rate from the perspective of the team, which was mainly analyzed through statistical methods in the early days. For example, Rubin used statistical methods to predict the scores of each game in 1958 (Rubin, 1958); scholar Barry proposed the Choice Models to predict the division champion of the US Major League Baseball (Barry & Hartigan, 1993); in 1995, some scholars predicted the long term performance of the MLB team (Kaigh, 1995); James et al. used statistical methods to answer questions related to baseball games (James, Albert, & Stern, 2004); the scholars, Yang and Swartz, use a two-stage Bayesian model

to predict the outcome of a major league team (Yang & Swartz, 2004). Based on the popularity of computers and the speed of computing, some scholars have begun to use different methods to predict baseball games: Donaker uses machine learning to predict and analyze MLB teams (Donaker, 2005); domestic scholars Lai and Chang use Bayes approach to predict the winner of the US Major League Baseball (Lai & Chang, 2008); Stekler et al. conduct a full-scale discussion of sports prediction (Stekler, Sendor, & Verlander, 2010); even until recently, researches on predicting the wins and losses of MLB teams has continued (Miller, 2011).

The second category of research focuses on the players themselves, such as the skills or positioning. Domestic scholars Chin-Cheng Chen and others published a series of articles on this topic, including: analysis of technical differences between Chinese professional league winning and losing investment (Chen & Chen, 2009), multi-scale approach to pitching technology targeting for starter pitchers (Chen & Chen, 2009), analysis of the difference between the team and the championship team by benchmarking management (Chen, 2012).

In summary, there is still a lack of research on the prediction of starting and bench ability of specific defensive positions. Both the pellet and the player urgently need reliable training objectives to improve the quality of players. For the player, the importance of starting is even higher than winning or not. Exploring the main difference between the bench and the starting opens up a new topic of sports prediction - first feature selection.

2.2. Feature Selection

The rise of feature selection is due to the rapid increase in the amount of data. In the face of a large number of high-dimensional data for data mining or computing, it is often prone to poor performance. In order to reduce or remove the data dimension and noise, it has begun to pay attention to it. The crux of the problem is to find representative and highly explanatory attribute variables (features) from the selected data set (several data and each data contains multiple attribute variables). Most classification problems belong to supervised learning, that is, the classification problem itself contains category attributes. When selecting features for classification problems of such supervised learning, it is necessary to measure not only the correlation between feature sets and categories, but also the correlation between features and features. Based on these two correlation measurements, the feature selection method is roughly divided into two different modes: Filter and Wrapper (Kohavi & John, 1997). The following is a description of Filter and Wrapper (Hung & Chang, 2009).

Most of the Filter feature selection methods measure the importance of a single feature, and regard the more important features as having sufficient information and could be classified and identified. This method could only filter features and combine them into feature sets, but cannot obtain the classification correct rate of the feature set. However, as a pre-processing technique before classification, it is necessary to combine the appropriate classification method with the evaluation method to obtain the classification correct rate. The features and categories selected by the Filter feature

selection method are related, but the correlation between features and features is completely ignored.

The Wrapper feature selection method is to optimize the feature set by continuously adding or deleting features, and using the target function in the learning algorithm to evaluate the feature set, and then using the search strategy to obtain the optimized feature set. This optimized search method may not be of high importance for a single feature dimension, but for the entire feature set, not only the feature set and the category are related, but also the features and features are related to each other. In general, if the classification problem has n features, a combination of 2n feature sets will be generated, and the best feature set must try all possible situations, thus requiring a lot of computation time (Kohavi & John, 1997). This study uses the Wrapper type feature selection method.

2.3. Classifier

The following two mainstream classifiers are introduced: Support Vector Machine (SVM) and Artificial Neural Network (ANN).

The SVM was brought up in 1995 by AT&T Bell Labs laboratory staff Vapnik and Cortes, and was developed from statistical learning theory based on the structural risk minimization principle (Cortes & Vapnik, 1995). The basic concept is a linear binary classifier that uses a linear separable hyper-plane as a classifier, which Vapnik defines as two sets of pre-labeled categorical values (1 or -1). The data, input linear function (linear function) for continuous training, and finally train to the best decision function of the two sets of data, thus forming the largest maximum margin between the two sets of data in the hyperplane classification, so that it divides the two groups of data to the most open, thus producing better promotion performance and higher classification accuracy. When the problem is linearly indivisible, the SVM could map the samples in the original space into the high-dimensional feature space, so that the original linear indivisible problem is transformed into a linearly separable problem (Hung & Chang, 2009).

ANN is a multi-layer perceptron (MLP) that mimics the principles of axons, dendrites, and synapse composed of nerve cells, which are then received by outside information. A computing network that stores, learns, and reacts, with a weighted function as a starting point for the forward-facing network; it can be divided into a supervised network such as MLP, an unsupervised network such as SOM, or a reinforced network such as a genetic algorithm (Li & Ku, 2010).

3. Research Methods

3.1. Research Process

Data collection is performed according to the appropriate defend position of the observation object, and then analyzed by linear regression. The key capabilities are selected according to the analysis results, and the key factors are selected to establish and predict the model, including various attack ability indicators and defensive

indicators as independent variables, and starting and bench ratio as dependent variables. Finally, collect suggestions of experts and commentators and conduct a difference analysis based on the results of the research. It can be explored whether the actual situation is consistent, and the pellets or players could be used as the basis for reinforcement based on the main features of starting.

The feature selection in the above process has been mentioned in the "Literature Discussion". The machine learning-based feature selection method in data exploration has been proved to be better than the traditional statistical method (Lin & Chen, 2008). It could be unrestricted from samples and dimensions, such as high-dimensional or low-sample factors, while traditional statistical methods are the opposite (Li & Yao, 2006). Meanwhile, data exploration and machine learning are good at dealing with nonlinear mathematical problems, so they could provide higher accuracy for complex data (Li & Ku, 2010). The steps for feature selection are as follows:

a) M * N Prepare the data set: When the data set contains M data and each data has N attribute variables, the data set matrix is expressed as M * N;
b) This data set is input to a specific feature selection algorithm for analysis, such as gene algorithm; this study uses WEKA software for algorithm analysis (Witten, et al., 1999), and parameter settings are directly use built-in default WEKA;
c) Feature selection algorithm outputs representative attribute variables;
d) The required training data is generated using three well-known feature selection methods, including gene algorithm, decision tree, and linear regression.

3.2. The Training and Testing of Predictive Model

The following is a comparison of the representative variables and the prediction performance outputted by the three feature selection algorithms. The representative variable explanatory ability selected by the algorithm is the prediction performance. The steps are as follows:

a) Prepare the data set
    a. Three same amount of data with different data sets, represented by Dataset-GA (Genetic Algorithm), Dataset-DT (Decision Tree), and Dataset-LR (Linear Regression), are representative variables selected by the feature selection algorithm.
    b. For each data set, classify each of its data as a starting or non-starting player (represented as category 1 and 2)
    c. Each data set is input into two different classification prediction algorithms (classifiers): SVM supports vector machine and ANN neural network, both of which are trained and tested by 5-fold cross validation for predictive model (Kohavi, 1995).
b) Each predictive model outputs an average prediction accuracy rate, that is, each data set has a prediction accuracy rate of SVM, DT, and ANN.

4. Conclusion and Discussion

4.1. Experimental Data Distribution

The information is obtained from the US Major League Baseball website (http://mlb.mlb.com/home) of 2015, and the training data is the general season results of the second baseman and shortstop of the League of Nations. The aforementioned "research scope" has already explained the principle of distinction between starting and bench, that is, if the actual number of starting of a certain player is 40 and games of the affiliated team played is 100 (the starting rate is 0.4), this experiment would consider such player is a bench player; those with a starting percentage greater than 0.5 are marked as starting players; if the starting percentage is lower than 0.5, but the player on defend position, who is the one with the highest starting percentage of the year, is also considered to be the starting player. Actual examine status of data distribution, 80~90% of the starting players' starting percentage is mostly exceeds 0.65, except for a few due to the fact that the season data is too little to be judged; and compared to the sports news of the season one by one, it could be confirmed that by the defined threshold 0.5 has reasonableness and credibility.

The nature of the shortstop and the second baseman is similar. The fielding positions of these two could be exchanged during the dispatch. Therefore, the general season data of all MLB shortstops and second basemen in 2009 would be combined into bench as test data. Table 1 lists the numerical value of starting and bench data in the general season of 2014 and 2015. It could be seen that there is not much difference between the number of starting and bench data. Such data distribution could enhance the fairness of the experiment and reduce the bias of the forecast, and it could also avoid predicting only one category and getting better classification results.

Table 1: Data distribution of second baseman and shortstop during 2015 and 2014 seasons

General season	Second baseman		Shortstop	
	Starting	Bench	Starting	Bench
2014 (training)	20	30	15	20
2015 (testing)	32	32	30	32

4.2. Feature Selection and Representative Ability Test

Table 2 is the selection result of three different feature selection methods. The focus of the fielding positions is the second baseman and the shortstop, which shows that there are key features of the second baseman of the League of Nations: on base capacity (H, 3B, HR, BB, 2B, OBP), speed (SB) and ball selection ability (SO, BB). Among them, except the good condition of on base and speed, the second baseman could not have too many strikes. For the League of Nations team's shortstops, comparing with second baseman, fielding percentage (FPCT) is an obvious and key starting factor. Different from the general public experience, the results of this experiment show that: the second baseman of the League of Nations is biased towards attacking characteristics, while the shortstop is focused on defensive ability.

Table 2: Feature selection results of various feature selection methods

Methods	Second baseman	Shortstop
Gene algorithm: Dataset-GA	H, 3B, SO, SB	H, BB, FPCT
Decision tree: Dataset-DT	HR, R, SLG	H, 2B, SB
Linear regression: Dataset-LR	R, H, HR, SO, SB, CS	H, HR, BB, SB, OBP, FPCT

The key features selected in Table 2 are further trained by the classifier, and the representative ability of the feature is determined by the superiority and inferiority of the classification ability. According to the above five cross-validation methods, the results are summarized in Table 3, and the following are the disclosed messages: (1) In general, the features selected by the gene algorithm are more representative; (2) The classifier SVM provides better training methods and prediction models, and the classification results are better than those of the ANN; (3) Comparison Table 2 and the gene algorithm in Table 3, the same conclusion could be made: the second baseman is attack-oriented, and the shortstop is more defensive.

Table 3: Classification results of various feature selection methods

Data set	Classification	Second baseman (5-fold)	Shortstop (5-fold)
Gene algorithm: Dataset-GA	SVM	**88%**	**94%**
	ANN	**83%**	73%
Decision tree: Dataset-DT	SVM	84%	89%
	ANN	77%	**80%**
Linear regression: Dataset-LR	SVM	86%	**94%**
	ANN	73%	**80%**

Finally, the interpretation of the applicability of the feature ability and the difference between the infielders is used. The data of the second baseman and the shortstop in 2009 is applied to the prediction model of Table 3, and the results in Table 4 are displayed as follows: The features extracted by this method have the ability to be inferred to other years, but there are significant differences between the use of the second baseman to predict the shortstop or the use of the shortstop to predict the second baseman.

Table 4: Predictive accuracy

Data set	Classification	2014 Second baseman (Training materials)		2014 Shortstop (Training materials)	
		2015 Second baseman	2009 Shortstop	2015 Shortstop	2009 Second baseman
Gene algorithm:		H, 3B, SO, SB		H, BB, FPCT	
Dataset-GA	SVM	**91%**	83%	87%	91%
Decision tree:		R, HR, SLG		H, 2B, SB	
Dataset-DT	SVM	**91%**	87%	**93%**	91%
Linear regression:		R, H, HR, SO, SB, CS		H, HR, BB, SB, OBP, FPCT	
dataset-LR	SVM	**91%**	83%	87%	91%

5. Conclusion

This paper is combined with feature selection methods and data exploration techniques, exploring the differences between the US Major League Baseball starting and bench players. The screening data obtained by the model could not only show high prediction accuracy, but also extend to the general season data of other years. The number of samples could be used regardless of the number of samples. In the difference between the ability of starting and the bench's field ability, the data exploration provides more objective information, and at the same time achieves 80~90% representation, and the key features are compared with the case data to prove the reliability of the study.

Finally, it is emphasized that the model established is generally effective for the general season, but for players who have not yet been promoted to the major league level, it is impossible to obtain information to predict the player. For the players who are about to be promoted, they would be recruited to the major leagues for spring training, but during the general season after the spring training, these players do not necessarily have the opportunity to be assigned as starting players. If we could use the spring training or the small league's data, we could predict the ability to be selected as a key indicator of the major league players, so as to play a predictive benefit and verify the correlation between the data.

References

Barry D., Hartigan J.A., 1993. Choice Models for Predicting Divisional Winners in Major League Baseball. Journal of the American Statistical Association 88(423): 766-774. doi:10.2307 / 2290761

Chen C.C., 2012. Establishing Quality Start Model of Chinese Professional Baseball League Using Logistic Regression. Journal of Physical Education Fu Jen Catholic University 11: 18-34. doi:10.29697/ JPE.201205.0002

Chen C.C., Cheng C.C., Chen T.T., 2005. Application of Standards Management Method in Professional Baseball Analysis—The 14th Annual Season of Chinese Professional Baseball League. Journal of Physical Education Fu Jen Catholic University 4: 206-218. doi:10.29697/ JPE.200505.0015

Chen C.C., Chen T.T., 2009. Starting Pitchers' Pitch Skills Positioning of CPBL in 2008-Multidimentional Scaling Analysis. Journal of Physical Education Fu Jen Catholic University 8: 109-125. doi:10.29697/ JPE.200905.0008

Chen C.C., Chen T.T., Chen Y.C., Yu F.H., 2001. A Study on the Differences between the Winning Pitchers and the Beating Pitchers in the Chinese Major League Baseball. Journal of Tamkang Sports 12: 240-247. doi:10.6976/ TJP.200912.0240

Cortes C., Vapnik V., 1995. Support vector networks. Machine Learning 20: 273-297.

Donaker G., 2005. Applying Machine Learning to MLB Prediction & Analysis. CS229 – Stanford University.

Feng J.H., 2010. A study on the winning factors in professional baseball games - an application of data mining technology. The 6th knowledge community seminar, Taipei City, Taiwan.

Hung Y.H., Chang P.J., 2009. Applying IG feature selection to improve SVM multi-category classification performance. The 17th symposium on fuzzy theory and its applications, Kaohsiung City, Taiwan.

James B., Albert J., Stern H.S., 1993. Answering Questions about Baseball Using Statistics. Chance 6(2): 17-30. doi:10.1080 / 09332480.1993.10542357

Kaigh W.D., 1995. Forecasting Baseball Games. Chance 8(2): 33-37. doi:10.1080 / 09332480.1995.10542458

Kohavi R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence 2: 1137-1143.

Kohavi R., John G.H., 1997. Wrappers for Feature Subset Selection. Artificial Intelligence 97(12): 273-324. doi: 10.1016/S0004-3702(97)00043-X

Lai Y.T., Chang C.L., 2008. Using bayes approach to forecast the winner of professional games—the case of professional baseball in Taiwan. PhD Thesis, Aletheia University, New Taipei City, Taiwan.

Li C.H., Ku C.J., 2010. Study on Application of Neural Network and Data Mining Techniques for Medical Diagnosis. Engineering Science and Education Journal 7: 154-169. doi:10.6451/JETE.201003.0154

Li C.H., Wu K.C., Hung C.H., 2006. Particle population optimization for feature selection and support vector machine optimization. The 11th artificial intelligence and application seminar, Kaohsiung City, Taiwan.

Li W.P., Yao C.C., 2006. A Research of Data Mining Applied to the Predictive Model of Fatty Liver. Master's Thesis, Chung Yuan Christian University, Taoyuan City, Taiwan. Retrieved from http://www.nownews.com/2011/06/17/732-2720983.htm#ixzz1PYKp4h2L

Lin J.H., Chen Y.C., 2008. Constructing an integrated credit rating model by using data exploration technology. The 2008 innovation management and new vision seminar, Kaohsiung City, Taiwan.

Miller K., 2011. Predicting Wins for Baseball Games. St. Lawrence University, Department of Mathematics, Computer Science and Statistics.

Rubin E, 1958. An Analysis of Baseball Scores by Innings. The American Statistician 12(2): 2l-22. doi:10.1080/00031305.1958.10481766

Stekler H.O., Sendor D., Verlander R., 2010. Issues in Sports Forecasting. International Journal of Forecasting 26(3): 606–621. doi:10.1016/j.ijforecast.2010.01.003

Witten I.H., Frank E., Trigg L., Hall M., Holmes G., Cunningham S.J., 1999. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems 192–196.

Yang C.H., Tu C.J., Wu K.C., Chang C.Y., Liu H.H., 2006. Tabu-PSO for feature selection. The 5th outlying islands information technology and application seminar, Kinmen, Taiwan.

Yang T.Y., Swartz T., 2004. A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball. Journal of Data Science 2(1): 61-73. doi:10.6339/JDS.2004.02(1).142