Forecasting MLB Playoff Teams Using GA-SVM

Ting-Chun Yu and Jui-Chung Hung*

Department of Computer Science, University of Taipei
Taipei 100, Taiwan
*juichung@utaipei.edu.tw

Abstract

This paper studies the Major-League Baseball (MLB) Playoffs forecast. In general, the MLB related records are so numerous and complex, so it is very difficult to read and do the useful prediction. And invalid records will also affect the accuracy, computing time, performance of the classifier.

Therefore, we proposed the method and named Genetic Algorithm - Support Vector Machine (GA-SVM) as the prediction. Generally, Genetic algorithm (GA) is the common filtering method, it can filter to remove the invalid features effectively, and the remain of the valuable feature can be used to predict the teams of the playoffs. The Support Vector Machine can also help to perform the classification of the feature filtering. As mentioned above, we combined the advantages of the GA and Support vector machine (SVM) to execute the prediction; it can avoid the over-fitting, local optimization and help to do the classification. We can also use the GA-SVM to distinguish whether the MLB teams have entered the playoffs or not.

For GA-SVM, we try to collect all the baseball team's batting, pitching and fielding records in 1995~2016 and to establish the useful classification of the model. Finally, we will compare the performance of GA-SVM with SVM and C4.5 methods.

Keywords: Major-League Baseball (MLB), Genetic Algorithm (GA), Support Vector Machine (SVM)

I. Introduction

Major League Baseball is the world's highest standard baseball league and is also one of the four major professional sports in North America. Because it is the highest-level baseball game, so loved by the world's baseball of fans. Per that the Forbes reports that 2016 MLB overall revenue is close to 10 billion dollars [1]. From this information, we can see that the MLB games caused so many business opportunities; it is included tickets, advertising, and commodity.

Before the MLB season kicks off, a lot of the commentator, fanatic and popular sports program are keen to predict this season of the most popular playoff team. Although the player's various data and the formula is so complex, but the researchers still hope the baseball records can be regularity and to establish a set of effective forecasting methods, the pellets and fans can refer to the method and to predict the playoff team.

Because the baseball games are the team sport, it cannot be decided by the one or two superstars to decide the outcome of the winning or losing game. Otherwise, every player has their own position, so we can divide the baseball records into the three categories: Batting, Pitching and Defense. As just mentioned, finding the reliable prediction with these data is one of the greatest challenges faces by fans today.

To be able to effectively predict the playoff team, this paper proposes a method in which the MLB records are filtered by the GA and then combine the robust classification of SVM to forecast the playoff teams.

Evolutionary Algorithm is very important in the field of the machine learning. It often used to search for the optimal solution, searching problems and feature detection. GA based on Darwinian's the theory of "survival of the fittest", and have been considered to solve the evolving problem most efficiently.

Genetic Algorithm was proposed by Holland in 1975 [2], this method to emulate the process of the natural evolution to achieve the final optimization and inherits the principle of evolutionary computation, which includes the steps in selection, crossover, and mutation.

Vapnik proposed SVM as a new method of pattern recognition since 1995 [3]. SVM is a supervised learning model for analyzing data in classification and machine learning algorithms. The advantage of SVM that has complete Statistical learning theory and data sets can be analyzed despite any forms of objects.

II. Proposed Method

2.1 Collect the MLB records

Up till today with the development of baseball game, the baseball records can be divided into the three parts: batting, pitching and fielding. This paper involved the baseball records from the baseball-reference website [4].

We use these baseball records as the individual features and then we will filter out the remain useful features. In this paper, total 59 baseball records are adopted and listed in TABLE I.

TABLE I. THREE CATEGORIES OF THE BASEBALL RECORDS.

Batting	R/G, R, H, 2B, 3B, HR, RBI, BB, SO, BA, BP, SLG, OPS, TB, GDP, HBP, SH, SF, BAbip, LOB, SB, CS
Pitching	RA/G, W-L%, ERA, SHO, SV, R, ER, BB, IBB, ROE, SO, HBP, BK, WP, BF, WHIP, SO9, SO/W, H, 2B, 3B, HR, BA, OBP, SLG, OPS, GDP, HBP, LOB, FIP
Fielding	DefEf, Ch, PO, A, E, DP, Fld%

MLB regular games usually will be end in October, In TABLE II, we could know the evolution of the playoff system. The 1995-2016 playoff system defines that American League and National League total have the 8 or 10 teams could be entered the playoff season.

 ISBN 978-1-5090-4897-7

TABLE II. THE REGULAR AND PLAYOFF OF THE SYSTEM.

	League	Regular Teams	Playoff Teams
1995-1997	AL	14	8
	NL	14	
1998-2011	AL	15	8
	NL	15	
2012-2016	AL	15	10
	NL	15	

By the above rule, we can set the label to distinguish which teams enter the playoff or not for these 21 seasons.

2.2 Data Normalization

Because the baseball records have their own meaning. Before the GA-SVM, we need to eliminate the disparity between the individual baseball records.

We adopted the Max-Min Normalization to eliminate the disparity for all collected baseball records [5].

Suppose that the total features of the baseball records are B and the numbers of the records, f_1, f_2,…, f_n.

And max_B and min_B are the extreme values of the collected baseball records. If the calculated value is f_j, f_j' is the calculated value after the normalization, the new range is $[newminB, newmaxB]$.

$$f_j' = \frac{f_j - min_B}{max_B - min_B}(newmaxB - newminB) + newminB \quad (1)$$

2.3 GA filter the valuable features

This paper adopts the GA to filter the useful baseball records. The GA executes the initial generation and the population is the set of the selected baseball records.

We use the binary encoding to encode for each baseball records and it means the single selected baseball record will be labeled as 1, if not used and will be labeled as 0.

GA will execute the fitness function for every generation and sorting the individual generation in accordance with the fitness function [7]. If the fitness function is higher than the other generation, it means the higher chance of being chosen. We adopted the fitness function as follows:

$$Fitness\ Function = \exp\left(\frac{\frac{\sum_{j=1}^{C_n} P_j B_j}{\sum_{j=1}^{C_n} P_j}}{\frac{C_s}{C_n}}\right) + \exp(-E * Z_E) \quad (2)$$

Total numbers of the baseball records are the C_n, the weight of the jth baseball record is P_j and B_j is the jth baseball record is selected or not (if the feature is elected stands for 1, if not selected is 0), C_s is the total numbers of the selected features, E is the classification accuracy rate of the training data and Z_E is the scaling factor of the E.

$$P_j = \left\|\overline{Y}_j - \overline{N}_j\right\|,\ j = 1,2,\dots,C_n \quad (3)$$

$\overline{Y}_j$ is the average of the jth baseball record of the playoff teams, $\overline{N}_j$ is the average of the jth baseball record of the non-playoff teams, P_j is the absolute norm of the $\overline{Y}_j$ minus $\overline{N}_j$, and it also means the weight of the jth baseball record, and then we can count the weight of every signal baseball record, as mentioned above if we always to choose the highest fitness, it may be happened fast convergence to local optimal solution and not the global optimal. We can adopt some method to solve this problem. Crossover is the process to create the excellent chromosome. Higher crossover rate it means the more pairs of the parent will have the next new generation and then the new generation will instead of the old generation. Mutation process is random to select the bit string and change to the bit information and to avoid the local optimum.

2.4 SVM Classification

After the above method, we get the initial generation and perform the SVM classification to get the accuracy rate.

SVM is a linear classification and the basic concept is to separate the data sets with a hyperplane. This hyperplane can determine which data points is closest and called Support Vector (SV), and the rest of the points do not need to do the calculation.

In the case of two-dimensional on a plane, we want to find a line that separates two different classifications and the boundary between these two sets is largest. Fig. 1 shows the SVM to find the separation hyperplane and support hyperplane. It can separate the circle and square samples with the maximum margin. Can also see that the support vector is closet the hyperplane.

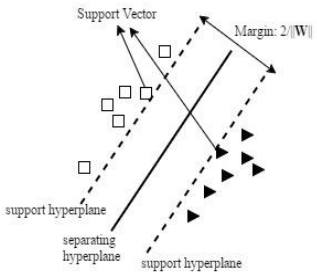

Fig. 1. Hyperplane with SVM

We suppose that the classifier need to map the data to the two dimension of the space and got the training set of N vectors $\{x_j, y_j\}_{j=1}^{N}$, where $x_j \in R^n$ is the input of jth eigenvector and $y_j \in \{+1, -1\}$ is an indicator vector and can also be called as label. And $\mathbf{W}$ is a weight vector and perpendicular to the hyperplane. Suppose to found one separating hyperplane to get the maximum margin with the two-support hyperplane. The definition of a linear equation [6], $f(\mathrm{x}) = \mathbf{W}^T\boldsymbol{X}_j - b$, $f(\mathrm{x}) > 0$, and separate $y_j = 1$ and $y_j = -1$,

$$\begin{aligned} \mathbf{W}^T\boldsymbol{X}_j - b \le -1, \forall y_j = +1 \\ \mathbf{W}^T\boldsymbol{X}_j - b \ge +1, \forall y_j = -1 \end{aligned} \quad (4)$$

and equal to

$$y_j(\mathbf{W}^T\mathbf{X}_j - b) - 1 \geq 0 \quad (5)$$

Because we want to get the maximum margin with the two-support hyperplane and the distance should be as $\frac{2}{\|\mathbf{W}\|}$ and equal to minimum margin $\frac{1}{2}\mathbf{W}^T\mathbf{W}$, and can be summarized as follows:

$$\begin{aligned} &\text{minimun margin} && \tfrac{1}{2}\mathbf{W}^T\mathbf{W} \\ &\text{subject to} && y_j(\mathbf{W}^T\mathbf{X}_j - b) - 1 \geq 0 \end{aligned} \quad (6)$$

In Eq. (6), we can found that all training data will fall within the limitation of the support hyperplane. and this equation also is the major problem in SVM.

Based on the above method, we created the GA-SVM process in Fig. 2.

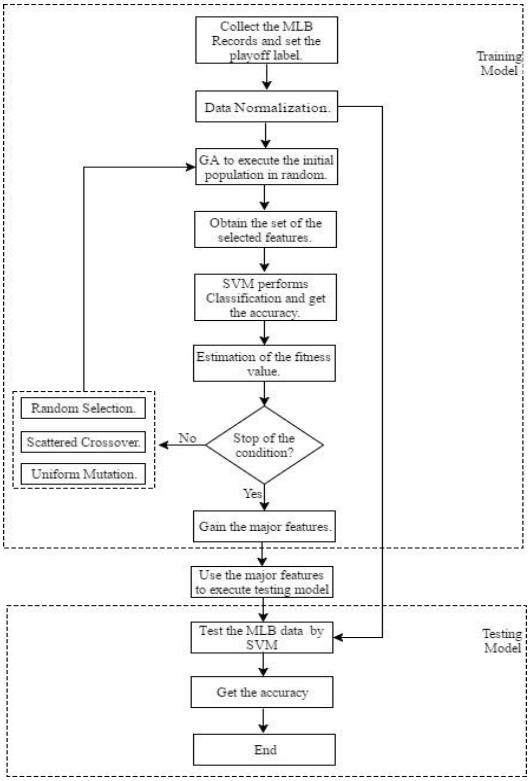

Fig. 2. FLOW CHART FOR THE GA-SVM

IV. Simulation Result

We conducted the experiment with the collected MLB data and total 59 baseball records and 654 team statics will be used. Otherwise, we suppose to use the LIBSVM in MATLAB and then do the classification by SVM and filter the useful features by GA [8]. From the above steps, total 5 features will be used and do the comparison with the C4. 5 and SVM.

C4. 5 is the commonly method of the machine learning [9], and we set the test option is Cross-validation Fold-4 and SVM will set and percentage split is 66% for the training data, and the remain 34% will test for the testing data.

The first comparison table is the C4. 5 comparison result, and shows in the table III as follows:

TABLE III
GA-SVM TO COMPARE WITH THE C4. 5 in 4-FOLD

GA-SVM		C4.5
Features Used	Accuracy	Accuracy
5	90.67	89.6

The second comparison table is SVM comparison result, and shows in the table IV as follows:

TABLE IV
GA-SVM TO COMPARE WITH SVM IN PERCENTAGE SPLIT-66%

GA-SVM		SVM
Features Used	Accuracy	Accuracy
5	92.34	88.34

V. Conclusion

In this paper, we proposed the GA-SVM method to do the prediction for MLB playoff teams. The comparing result has higher accuracy than the C4. 5 and traditional SVM. Therefore, this proposed method could provide the prediction in the future.

Acknowledge

This research was supported by the Ministry of Science and Technology, R.O.C. Project number: MOST 104-2221-E-845-001-MY2

References

[1] Forbes, https://www.forbes.com/sites/maurybrown/2016/12/05/mlb-sees-record-revenues-approaching-10-billion-for-2016/#652a36227088
[2] JH Holland, “Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence (3rd Edition.)”, MIT Press, Cambridge, MA, 1994.
[3] CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. *Machine learning*, 1995, 20.3: 273-297
[4] Baseball-Reference, http://www.baseball-reference.com
[5] Kamber, Micheline, Jiawei Han, and Jian Pei. “Data mining: Concepts and techniques”. Elsevier, pp. 113-114, 2012.
[6] SCHOLKOPF, Bernhard; SMOLA, Alexander J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, pp.9-16, 2001.
[7] M. H. Zhong, J. C. Hung, Y. C. Yang, C. P. Huang, “GA-SVM classifying method applied to dynamic evaluation of taekwondo” International Conference on Advanced Materials for Science and Engineering, pp. 534-537, 2016.
[8] C.C. Chang, C.J. Lin, “LIBSVM: A library for support vector machines”, *ACM Transactions on* Intelligent *Systems and Technology*, Vol. 2, No. 27, 2011.
[9] QUINLAN, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

 ISBN 978-1-5090-4897-7