Prediction of Football Match Results Based on Model Fusion

Quan Zhang[1,2]
[1]Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang, China
[2]School of Information Engineering, East China University of Technology, Nanchang, China
280482028@qq.com

HongZhen Xu[1,2*]
[1]Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang, China
[2]School of Information Engineering, East China University of Technology, Nanchang, China
xuhz@ecut.cn

Li Wei
School of Information Engineering, East China University of Technology, Nanchang, China
1337520991@qq.com

LiangQi Zhou
School of Information Engineering, East China University of Technology, Nanchang, China
751966851@qq.com

ABSTRACT

In recent years, the prediction of the results of competitive sports events has received wide attention. For example, in the football field, the prediction results have been used in team management, analysis and training. Therefore, the prediction of football match results has research significance and value. However, the existing prediction models of football match have problems of poor generalization ability and low accuracy. In response to these problems, this paper proposes a football match result prediction method based on model fusion. The paper takes the Chinese super league as the research object. Firstly, we obtain the data of the Chinese super league in 2013-2018 from the football association's website. Then, based on the data analysis, according to the winning factors of the football match, the characteristics are selected, three machine learning methods which are Support Vector Machine(SVM), random forest and Bayesian, are used as a primary classifier to separately train the data. At last, SoftMax is used as a secondary classifier to establish a prediction model. By comparing the training results of a single classifier, our method can achieve better prediction accuracy.

CCS Concepts

• **Computing methodologies → Ensemble methods**

Keywords

football prediction; machine learning; SVM; model fusion

ICIAI 2019, March 15–18, 2019, Suzhou, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6128-6/19/03…$15.00
https://doi.org/10.1145/3319921.3319969

1. INTRODUCTION

Football is one of the most influential and highly regarded sports. It is also one of the most economically valuable projects in the sports field. The annual GDP of the football industry is worth 500 billion US dollars. It is known as the “17th largest economy”. Such huge commercial value makes the prediction of football matches more important. For the team, the prediction of the game results will help the team to adjust to a certain extent before the game and deploy more effective tactics. However, there is a saying on the court that "football is round, anything can happen." The prediction of match results is a difficult problem, and there are too many uncertain factors. For example, the advantages of home and away, weather, game time and referee scale, etc., the effect of traditional prediction method is not satisfactory. Therefore, how to effectively predict the results of the game has become a research hotspot.

This paper obtains data from the legal Chinese Super Football official website, combines the factors of football success with some attributes in the original data to construct a data set, and proposes a football game prediction method based on model fusion. Specifically, the primary classifier we use are three classifiers, SVM, random forest and Bayes, which are widely used in financial forecasting, event prediction, weather forecasting, and other forecasting tasks. The results of the three classifiers were then fitted using a SoftMax classifier.

The main contents of this paper are as follows: the related work is summarized in the second part, and the prediction model based on model fusion is introduced in the third part. The experimental results are given and discussed in the fourth part. The paper is summarized in the fifth part.

2. RELATED WORK

In the past, the forecasting methods mainly include the Bayesian model and the fuzzy comprehensive evaluation model, but they each have their own shortcomings. The Bayesian model is too focused on the historical victory or defeat, while the fuzzy comprehensive evaluation model is concerned with many winning factors, the latitude the data sets is too high, and the data acquisition in the early stage is also the main difficulty of this method. In recent years, with the rapid development of machine learning methods, it has become a trend to solve problems in the field of football based on machine learning methods. At present, other researchers have made some progress in the prediction of the event. Li ZhongXun[1] uses RNN to establish the football player scoring model, and then combines the model with logistic regression to predict the matches results, and gets a good result. Igiri et al.[2] used Logistic regression and ANN to analyze the performance of different teams, players and coaches in a particular season, however this analysis method is only effective for a specific season, not ideal for the new season; Kolbush et al.[3] used logistic regression and Markov to establish a football prediction model; Nilay Zaveri[4] uses Logistic Regression, Random Forest, ANN, Naive Bayes, and SVM to achieve predictions for La Liga respectively; Berrar et al.[5] integrate domain knowledge into machine learning methods to design predictive models, and adopts KNN design in the process of model design. then use XGBoost to fine-tune the model, and achieved good results; Hubáček et al. [6] used the gradient enhancement tree to predict the results of the football matches; in the past few years, the Bayesian network is considered to have a good application in the field of football prediction. Nazim[7] used Bayesian networks to achieve better results, however, he does not use a new season data as the test data.

Different from the above method, this paper first analyzes the winning factors of the football match[8], combines the historical data of the season, builds the data set, and then uses SVM, random forest and Bayesian to train separately, finally, based on the model fusion idea, SoftMax was used as the sub-classifier to learn their training results. After testing, compared with the methods in the above literature, this research has a better prediction effect in the new season data, and the whole prediction model reflects the better generalization ability.

3. PREDICTION MODEL

The prediction model of this paper is based on SVM, random forest and Bayesian algorithm, to construct a single model respectively, and then use SoftMax as a secondary classifier to fuse the above three models to build a fusion model, and compare the prediction performance of each model. The three selected single models have good diversity, less correlation, and similar performance, which meet the basic conditions of model fusion.

3.1 Get sample data

This article uses the China Football Association Super League 2013-2018 match data as a sample of research, the Chinese Super League is a total of 16 teams, each season to conduct 30 rounds of competition, we obtain sample data such as integral ranking data, single season shooter data, and the data of each game, etc., Table 1 is the 30th round of the 2018 season of the Chinese Super League part of the data, including the top 5 team in 2018.

This article uses a total of six seasons of information, a total of 240 matches each season, so there are a total of 1440 matches, we use the 2013-2016 data as a training set, 2017 and 2018 data as a verification set.

Table 1. Chinese super league partial points data

rank	team	win	draw	lose	goals scored	goals conceded	Goal difference	points
1	Shanghai SIPG	21	5	4	77	33	44	68
2	Guangzhou Evergrande Taobao	20	3	7	82	36	46	63
3	Shandong Luneng Taishan	17	7	6	57	39	18	58
4	Beijing Sinobo Guoan	15	8	7	64	45	19	53
5	Jiangsu Suning	13	9	8	48	33	15	48

3.2 Data set construction

After analyzing the original sample data, and based on the winning factors of football matches, this paper proposes 14 characteristics of the matches to construct the input data set of the prediction model.

The characteristics of the data set are shown in Table 2.

Output: The league match system allows for a draw, so the output is divided into 3 cases, that is, the home team wins, the home team loses, The home and guest team draw.

Table 2. Input dataset characteristics table

Feature	Instruction
Home Team	Defines the name of the home team
Away Team	Defines the name of the away team
Home Goal	The number of goals scored by the home team in this season
Home Lose	The number of lose scored by the home team in this season
Away Goal	The number of goals scored by the away team in this season

Away Lose	The number of lose scored by the away team in this season
Home Suspension	The number of players suspended from the home team this round
Away Suspension	The number of players suspended from the away team this round
Home Points	Home team points before the round
Away Points	Away team points before the round
Last Season Results	The results of the two teams in the home team's home match last season (win, draw or lose)
Home Evaluation	Based on home team value and head coach ranking
Away Evaluation	Based on away team value and head coach ranking
Match Time	Afternoon or evening

3.3 Prediction model based on SVM

SVM is a machine learning method based on statistical learning theory. It is not based on the least risky approach, but uses the least risk of structuring. Therefore, it has better generalization ability[9]. It shows a big advantage when solving small samples, nonlinearities and high-dimensional pattern recognition. Although there are many historical data in sports competitions, especially group sports, rely too much on numbers of historical information, and the effect of prediction model is not ideal. Therefore, it is appropriate to select the data of recent seasons from the training data used in the event prediction of football league. With less season information, there will be less overall training data. Therefore, SVM is adopted as the primary classifier.

The classical SVM method is a dichotomous classification algorithm. The most basic idea of classification is to find a dividing hyperplane in the sample space based on the training set, and separate the samples of different categories. If the feature data itself is difficult to separate, that is, there is no hyperplane in the sample space to separate the samples, then the appropriate surface can complete the classification task by mapping the sample data from the low-dimensional indivisible space to the high-dimensional separability space, and find the classification plane, As shown in the Figure 1.

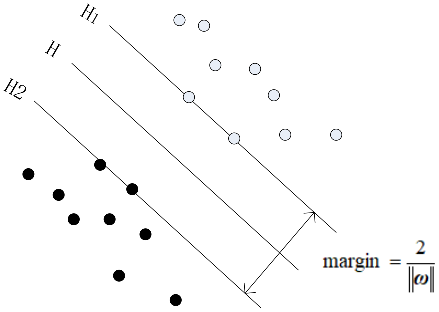

Figure 1. SVM optimal classification surface map

The original data points cannot be linearly distinguished. When they are transformed into higher-dimensional space, they can be classified by a plane. The specific method is to utilize a kernel function. The mapping transformation of feature vector ***x*** from low-dimensional space to high-dimensional space $\Phi : \mathrm{x} \rightarrow \mathrm{F}$can be obtained as Equation (1).

$$\mathrm{x} \rightarrow \Phi(x) = (\Phi_1(x), \Phi_2(x), \ldots, \Phi_l(x))^T \tag{1}$$

The hyperplane partitioning equation of SVM is as Equation (2).

$$\mathrm{y}(x) = w^T \Phi(x) + b \tag{2}$$

According to the partition equation of hyperplane, the decision equation of classification is as Equation (3).

$$\begin{aligned} \mathrm{y}(x_i) > 0 \Leftrightarrow y_i = +1 \\ y(x_i) < 0 \Leftrightarrow y_i = -1 \end{aligned} \tag{3}$$

Further infer the Equation (4).

$$\mathrm{y}_i . y(x_i) > 0 \tag{4}$$

Commonly used kernel functions are as follows: linear kernel function, polynomial kernel function, Gaussian kernel function (RBF), Laplacian kernel function and Sigmoid kernel function. The RBF kernel function is selected here to classify high dimensional samples. The Gaussian kernel function formula is as Equation (5).

$$K(x_i, x_j) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}), \quad （\sigma > 0） \tag{5}$$

Where σ is the bandwidth of the kernel function, the kernel function is used as a sample transformation formula, in the process of using, the formula is directly substituted into the expression $\Phi(x_i)$ in the formula (1). The parameter σ refers to the number of support vectors in the model. By adjusting this parameter, the model is optimized. When the value is large, the number of support vectors in the model will be more, and the model is easy to over-fitting. Otherwise, the number of support vectors will be less, and the model is simpler.

In this paper, it has been clarified that there are three kinds of output values, so this is a multi-classification problem. SVM is a classic dichotomous classifier, so the multi-classification SVM based on OVO SVMs (one-to-one method) will be used here: libSVM model. To establish the prediction model.

Start with three categories: W(win), D(draw), and L(lose). During training, libSVM selects the vector corresponding to (W, D), (W, L), (D, L) as the training set, and then obtains 3 training results. At the time of testing, the corresponding vectors were tested on three results.

3.4 Prediction model based on random forest

This paper chooses a random forest regression model in the primary classifier, which can process high-dimensional data without feature selection. In addition, it can give important characteristics after training. The generalization ability of the model is relatively strong, and the interaction between features can be detected during the training process.

The random forest is based on the integrated learning method[10,11], which uses the bagging method. It through the bootstrap resampling technology, T samples were randomly selected repeatedly from the original training sample set N, and a new training sample set was generated, then T classification trees were generated according to the self-service sample set to form the random forest, the classification result of the new data is determined according to the score formed by the classification tree voting. In the combination strategy, this paper adopts absolute majority voting methods, as shown in Equation 6.

$$H(x) = \begin{cases} c_j, & if \sum_{i=1}^{T} h_i^j(x) > 0.5 \sum_{k=1}^{N} \sum_{i=1}^{T} h_i^k(x); \\ reject, & otherise. \end{cases} \qquad (6)$$

3.5 Prediction model based on Naive Bayes

Naive Bayes is mainly based on the independent assumption of attribute conditions. In this paper, the input data has 14 attributes, and the output category has 3, so there are 14 feature items $(w_1, w_2, \cdots, w_n)$ in d. For a given class $c_k (k = 1, 2, \cdots, 3)$, the probability that d belongs to the class c_k is shown in Equation 7:

$$p(c_k \mid d) = Max\{p(c_1 \mid d), \; p(c_2 \mid d), \cdots p(c_n \mid d)\} \quad (7)$$

According to Bayesian probability formula, Equation 8 can be obtained:

$$p(c_k \mid d) = \frac{p(d \mid c_k) p(c_k)}{p(d)} \qquad (k = 1, 2, \cdots, 3) \qquad (8)$$

Where, Equation 9 is as follows:

$$p(d \mid c_k) = p(w_1, w_2, \cdots, w_n \mid c_k) \qquad (9)$$

The denominator $p(d)$ in Equation 8 is independent of the category, so it can be ignored when comparing the maximum values in Equation 7, therefore, it is only necessary to calculate the probabilities a $p(c_k)$ and $p(d \mid c_k)$ to classify d.

In Equation 8, $p(c_k)$ is a prior probability, which is easy to calculate, but the calculation of $p(d \mid c_k)$ is difficult, especially when the number of feature items is large and the degree of dependency between feature items is high, the calculation will be extremely time consuming. In order to simplify the calculation, a conditional probability independent hypothesis is introduced, which assumes that the features are independent of each other. This is the naive Bayesian filter[12,13,14], then Equation 9 can be converted to Equation 10:

$$p(d \mid c_k) = p(w_1, w_2, \cdots, w_n \mid c_k) = \prod_{i=1}^{n} p(w_i \mid c_k) \quad (10)$$

3.6 Model fusion

Model fusion has methods such as Stacking, Blending and Voting[15]. It is a powerful technique to reduce the generalization error by increasing the diversity of the algorithm, and finally achieve the improvement of the accuracy of the model. The model fusion has two basic elements: one is the correlation between them should be as small as possible, second, the performance between the single models is not much different. This paper adopts the Stacking method to design the prediction model. The basic idea of Stacking is to use large numbers of base classifiers, and then use another top-level classifier to fuse the prediction of the base classifier, aiming to reduce the generalization error. The secondary classifier used in this paper is multiple classification logistic regression, which is the SoftMax classifier. The whole model fusion process in this paper is shown in Figure 2:

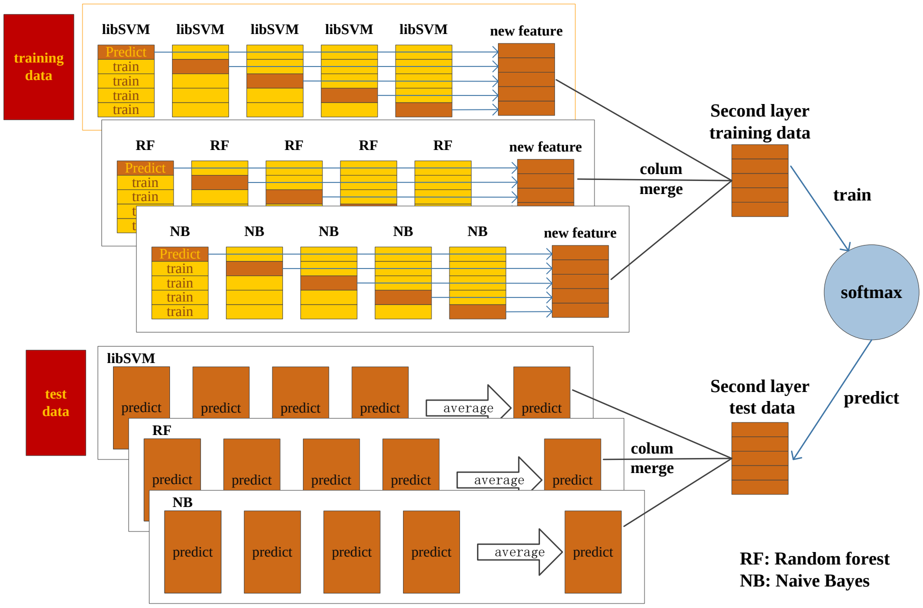

Figure 2. Model fusion process

4. EXPERIMENT RESULT

In this paper, the 2017 and 2018 Chinese Super League competition data were used for testing during the experimental test. We use SVM, RF, and Bayesian to establish a predictive model test to compare with the designed prediction model. In this model, SVM uses L_0 as a regular term, and the results are shown in Table 3. After that, we used the L_0, L_1 and L_2 regular terms in the SVM in the primary classifier to test the prediction model. The experimental results are shown in Table 4.

Table 3. Comparison of accuracy with different models

Dataset	SVM	RF	NB	Our model
2017	65.3±0.1	71.4±0.2	67.9±0.4	76.5±0.2
2018	63.7±0.3	70.5±0.3	64.1±0.1	75.7±0.4

Table 4. Comparison of accuracy with different regularization

Dataset	Our model+ L_0	Our model+ L_1	Our model+ L_2
2017	76.5±0.2	75.9±0.1	76.9±0.2
2018	75.7±0.4	73.5±0.2	76.1±0.1

It can be concluded from Table 3 and Table 4 that our model based on the Stacking method has achieved good results in testing the new season information, and the model works best when the SVM adopts L2 regularization.

5. CONCLUSION

This paper proposes a model of football match prediction based on model fusion. Based on the model constructed by SVM, random forest and Bayesian, the SoftMax classifier is used to perform the new data set constructed by the training results of the above three models. After training, the obtained prediction model has higher correct rate on the verification set than the single classifier, and achieves better results. The whole prediction model has strong generalization ability.

6. ACKNOWLEDGMENTS

This work was supported by the Program of Training Young Scientists of Jiangxi Province under the Grant No. 20142BCB23017, the Scientific Research Plan Projects of Education Department of Jiangxi Province of China under the Grant No. GJJ151538 and GJJ160554, and the Opening Project of Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology under the Grant No. JELRGBDT201802.

7. REFERENCES

[1] Li. Z. X. 2017. *Research on the Theory and Application by Applying Machine Learning to Player Ratings in Football Games*, University of Electronic Science and Technology of China.

[2] Igiri C P, Nwachukwu E O. 2014. *An improved prediction system for football a match result*. IOSR Journal of Engineering (IOSRJEN),4: 12-20.

[3] Kolbush J, Sokol J. 2017.*A Logistic Regression/Markov Chain Model for American College Football*. International Journal of Computer Science in Sport, 16(3): 185-196.

[4] Zaveri, N., Tiwari, S., Shinde, P., Shah, U., & Teli, L. K. 2018. *Prediction of Football Match Score and Decision Making Process*. International Journal on Recent and Innovation Trends in Computing and Communication, 6(2), 162-165.

[5] Berrar, D., Lopes, P., & Dubitzky, W. 2018. *Incorporating domain knowledge in machine learning for soccer outcome prediction*. Machine Learning, 1-30.

[6] Hubáček, O., Šourek, G., & Železn ý, F. 2018. *Learning to predict soccer results from relational data with gradient boosted trees*. Machine Learning, 1-19.

[7] Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. 2017. *Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)*. In IOP Conference Series: Materials Science and Engineering (Vol. 226, No. 1, p. 012099). IOP Publishing.

[8] Hou H. S,Mimadunzhu， Hou B,Guo J. L. 2017. *Key Winning Factors and Winning Formula of Football Games*. Journal of Beijing Sport University,40(11):105-110.

[9] Bai P.F, An Q, Nicolaas. F. R, Li. N, Zhou G. F. 2017. *Internet Credit Personal Credit Assessing Method Based on Multi －Model Ensemble*. Journal of south China Normal University(Natural Science Edition), 49(6):119-123.

[10] Xiao, Y., Wu, J., Lin, Z., & Zhao, X. 2018. *A deep learning-based multi-model ensemble method for cancer prediction*. Computer methods and programs in biomedicine, 153, 1-9.

[11] Xiao, J., Xiao, Z., Wang, D., Bai, J., Havyarimana, V., & Zeng, F. 2018. *Short-term traffic volume prediction by ensemble learning in concept drifting environments*. Knowledge-Based Systems.

[12] Zhang, W., Qi, Y., Zhou, Z., Biancardo, S. A., Shen, M., & Wang, Y. 2018. *A Method of Speed Data Fusion Based on Bayesian Combination Algorithm and Markov Model*. No. 18-02593.

[13] Chen, S. S., Cao, J. J., Gan, L. L., Song, Q. G., & Han, D. 2018. *Experimental study on generalization capability of extended naive Bayesian classifier*. International Journal of Machine Learning and Cybernetics, 1-15.

[14] Liu, Z., Pan, Q., Dezert, J., Han, J. W., & He, Y. 2018. *Classifier fusion with contextual reliability evaluation*. IEEE transactions on cybernetics, *48*(5), 1605-1618.

[15] Yin, Z., Zhao, M., Wang, Y., Yang, J., & Zhang, J. 2017. *Recognition of emotions using multimodal physiological signals and an ensemble deep learning model*. Computer methods and programs in biomedicine, 140, 93-110.