Prediction of Football Match Result using machine learning on Hive

Guang Yang
New York University
New York, U.S.A
gy552@nyu.edu

***Abstract*—Predictive systems have been employed in various areas to predict events and results, especially the football match, the most popular sports around the world. In this paper, we present a system predicting the result of football matches in the English Premier League (EPL) using big data tools and machine learning algorithms. This system takes advantages of big data, uses Apache Hive[1] and Hivemall[2] to process, and gives accurate prediction based on the features we designed.**

Keywords—Football, prediction, analytic, Hive, machine learning

I. INTRODUCTION

Football has always been a fantastic sports which attracts lots of people, prediction of football match result is the same. Researchers have spent many years finding the algorithm which may give the better prediction, and several methods have been tried. Traditional statistical method is widely used in 90s, but they yielded poor accuracy. In the past years, with the heat of machine learning, a lot of machine learning algorithms have been employed to predict the football match result, including Naive Bayes(NB), Support Vector Machine(SVM), even the Neural Network(NN). But under the lack of data and features, they didn't give a satisfying result. In this paper, we have proposed a new system which exploit the big data tools: Hive and Hivemall, to extract enough features from the data, and combined machine learning algorithms including Soft Confidence Weighting (SCW), Random Forest (RF), Adaptive regularization of weights (AROW), to perform a relatively comprehensive prediction of the football match result.

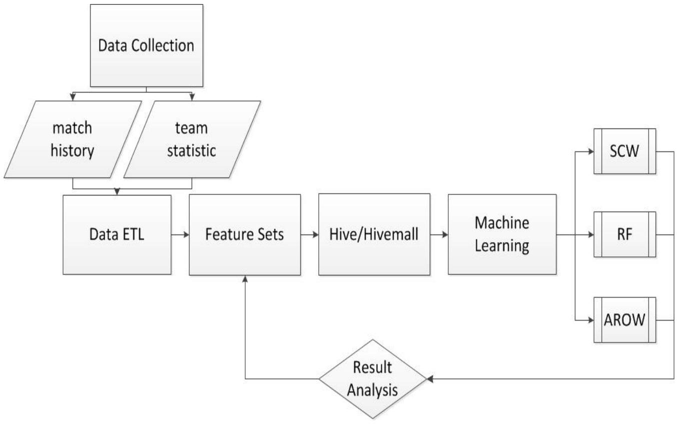
II. MOTIVATION

English Premier League (EPL) is the most popular worldwide watched league with a TV audience of 4.7 billion people. Many sports fans watched it every year, the results are what they talked and bet most. As a fan of football, digging out the relation between data and the match result will be the most fascinating thing. This analytic also has real world applications for gambling, coaching improvements and journalism. In this paper, we will show the methodology we used to predict the result based on the match history.

III. RELATED WORK

Lots of research on this topic have already been done. In recent years, D. Buursma has proposed a system using six different machine learning algorithms to predict the results[4]. Based on his discovery, I.C. Peace has raised a new system involving more features and more powerful machine learning algorithms like artificial neural network(ANN) and genetic algorithm to gain a better accuracy[3]. Besides the machine learning method, there are still many statistical method which reached satisfying performance. F. Louzada came up with a nearly pure mathematical algorithm to compute a rank for each team and gave the final rank to predict the result[6]. Mixing the machine learning and statistical method has been proved more effective. A. Gangal proposed a system implementing Naive Bayes, ANN and genetic algorithm to calculate a score, and using it to do the prediction[7]. Another score-based research used ELO rating as the basis, and worked really well[5].

IV. DESIGN



This flow chart has shown the system we proposed. We collected data from two websites using crawler, and then merged them into our database - Hive. We extracted features from the data, and created new tables for these features. Then we used these features to train machine learning models, and test several machine learning algorithms on them. Finally we test our model on data from one season, and analysis the performance of our feature sets.

We will show our whole system in three parts. In part A, we will discuss the crawler and data sources we used. In part B, we will list the features we extracted. And in part C, we will cover the algorithms we implemented.

A. Data Collection

We collected data from two data sources. The first is a github stored the match history of all past EPL matches. But the data is messy and need hard work to fix. An additional pre-processing using Map-Reduce is employed to normalize the data format. The second is a statistical summary of each team in each season. This webpage is dynamic generated through java script, so traditional crawler can't be used to extract the information from it. We exploited Htmlunit to simulate the browser client and run the java script to fetch the information from server, then used Jsoup for cleaning and regex matching. Combining these two parts of data, we can get a more comprehensive description of the teams, which will help us to find out the relation between data and match result much easier.

B. Features

We have extracted over twenty features from original data, some are listed below:

- The goals shot by home/away team in the last season
- The possession ratio of home/away team in the last season
- The success rate of passing of home/away team in the last season
- The goals scored by home/away team in the last season
- The goals conceded by home/away team in the last season
- The number of home matches won by home team in the last season
- The number of away matches won by away team in the last season
- The goal difference of home/away team in the last season

Because the EPL is a round match, in each season each team will play against every other team for two matches, as home and away team respectively. So in each season, 20 teams will perform 380 matches in total. We used only the data of last season to predict the outcome of the incoming season.

Several feature combinations have been tried and we found lots of interesting conclusions here. We will talk about the feature sets we selected in the results section later.

C. Algorithms

There are three main machine learning algorithms we implemented, namely the Adaptive regularization of weights (AROW), soft confidence weighting (SCW) and random forest (RF). There are still some other algorithms been tried, but they yielded disappointing result which we didn't include here. Then we will simply describe each algorithm respectively.

AROW[8]: Adaptive regularization of weights, which performs adaptive regularization of the prediction function upon seeing each new instance, allowing it to perform especially well in the presence of label noise. It helps to get rid of some non-separable data and deals well with the large margin training.

RF[9]: Random Forest, which is itself an ensemble learning method for classification and regression. It constructs a multitude of decision trees at training time and outputs the mode of the classes or mean of the classes predicted by each individual tree. This algorithm can correct the over-fitting problem which may be caused by pure decision tree algorithm.

SCW[10]: Soft Confidence Weighting, which is an update of the original famous online learning algorithm Confidence-Weighted algorithm. In Confidence Weighted learning algorithm, the classification model is drawn from a Gaussian distribution, which enables the model to be effectively updated with the second-order information of the data stream. Soft Confidence Weighting algorithm, in addition, handles some non-separable dataset and noisy dataset better.

V. Results

In the experiments, we split the whole dataset into two parts: the train data and the test data. 80% of the original data will be used as train data and 20% of them is the test data. We shuffled the data first to ensure that the labels are distributed in balance.

Then we tried three algorithms on lots of combinations of features. Features are grouped in home/away team pair, so each group contains two features for home team playing as home team and away team playing as away team respectively. Here we list some useful features below and give the test result of combinations of these features.

1. The possession ratio of home/away team in the last season.

2. The success rate of passing of home/away team in the last season.

3. The number of matches won by home/away team in the last season.

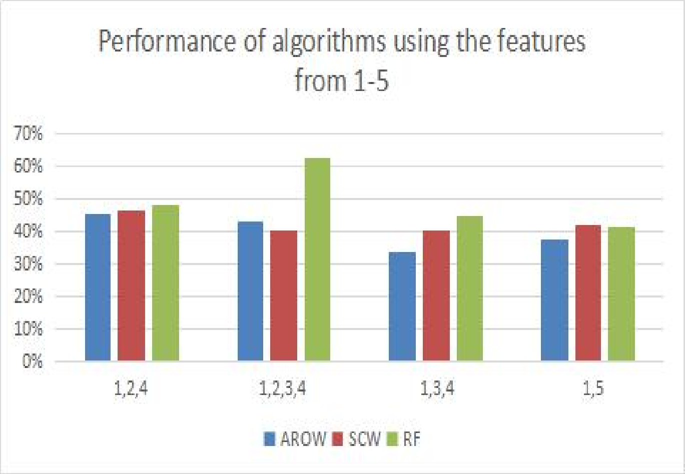
4. The goals scored by home/away team in the last season.

5. The goal difference of home/away team in the last season.

Table 1. Performance of algorithms using the features from 1-5

Algo/ Feature	1,2,4	1,2,3,4	1,3,4	1,4	1,5
AROW	45%	43.13%	33.4%	36.25%	37.12%
SCW	46.25%	40.31%	40.31%	50%	41.88%
RF	47.81%	**62.19%**	44.69%	46.88%	41.56%

Figure 1. Compare of algorithms using features from 1-5



From the table 1 above, we can see that the highest accuracy among these combinations of features are 45% (AROW) , 62.19% (RF) , 50% (SCW) respectively. And from the figure 1, we can found that AROW and SCW algorithm have similar accuracy, while RF usually has higher accuracy than them. Another insight is, RF algorithm has higher accuracy when many useful features are combined, but for AROW and SCW algorithm, too many features would lead to decrease of the accuracy.

There are also many other features which aren't so much important to the total performance or even make it worse. We list them below and give the test result of combinations of these features.

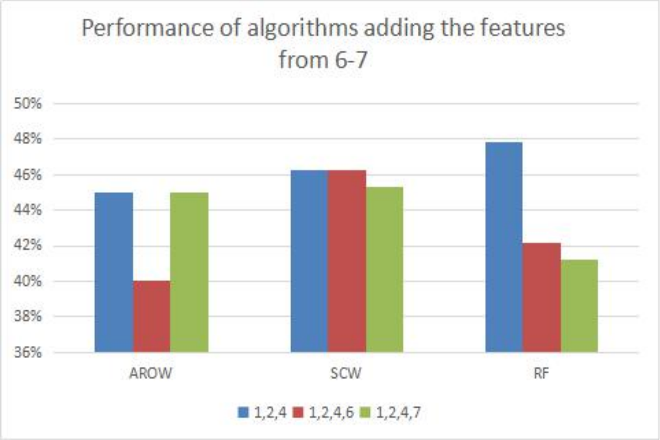
6. The goals conceded by home/away team in the last season.

7. The number of matches lost by home/away team in the last season.

Table 2. Performance of algorithms adding the features from 6-7

Algo/ Feature	1,2,4,6	1,2,4,7
AROW	40.06%	45%
SCW	46.25%	45.31%
RF	42.19%	41.25%

Figure 2. Compare of performances adding the features from 6-7



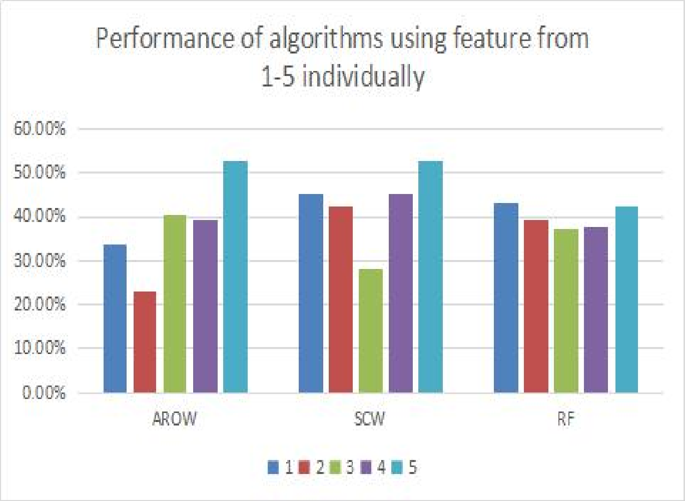
From the table 2 above, we can see that adding feature pair 6, 7 to the original feature combination 1,2,4 will not change the accuracy or even lower the accuracy. Among these three algorithms, SCW seems to be the most stable algorithm which can get rid of these noisy features, and due to the ensemble learning property, RF will be largely affected by some noisy or useless features. But once a feature is proved useful, adding it into RF will definitely increase the accuracy while SCW or AROW may not. Here we can see from figure 2 that the accuracy is decreased by more than 6% when add 6,7 features to RF algorithm, but it doesn't affect SCW or AROW algorithms too much.

We also test the individual performance of features from 1-5, and try to find the most important feature among these useful features.

Table 3. Performance of algorithms using feature from 1-5 individually

Algo/ Feature	1	2	3	4	5
AROW	33.75%	23.13%	40.31%	39.06%	**52.81%**
SCW	45%	42.5%	28.13%	45.31%	**52.5%**
RF	43.13%	39.38%	37.19%	37.81%	42.5%

Figure 3. Compare of performances using feature from 1-5 individually



From the table 3 and figure 3 above, we can see that AROW and SCW algorithm reach the highest accuracy when using only the feature pair 5 which is the goal difference of the home/away team. This performance is even better that all the performances when using combinations of feature pairs. But RF doesn't reach the highest accuracy with the single feature pair 5, which proves the assumption we made about RF algorithm that it need enough useful features to reach a high accuracy. Though it doesn't help all the algorithms reach their highest accuracy, the goal difference itself is the most important one among all the features. Beside the goal difference, we can see that the goal scored is also very important which helps all the algorithms reach the accuracy around 40%. Then goes the possession rate, which helps two of the algorithms reach the accuracy above 40% while AROW algorithm reaches only 33.75% with it.

From all the result shown above, we can draw several interesting conclusions about the features and the algorithms from this dataset.

For the features, we found that the goal difference or goals scored are essential to the result, which get 40%-50% accuracy even when used individually. Also, the possession rate of a team is important too. There are also some other useful features including the success rate of passing and the number of matches win. The remain features seem to provide little help to the prediction result, as they may be random over the matches or can be concluded from the other features.

For the algorithms, we found that SCW and AROW algorithms are relatively more stable, they have nearly the same accuracy. They have good noise resistance, but would perform worse when too much features are added. So they reach their highest accuracy on the single feature pair 5. In contrast, RF algorithm seems unstable in most of the time due to that it's a random algorithm. But there still some property we can conclude. RF algorithm would perform better when many useful features are added, that is, due to the ensemble property it has, the more relative features are added, the better performance it will have. But it has low noise resistance, so its accuracy will decrease dramatically when adding some useless features.

Another finding is that these three algorithms respond different to the same feature. That is, the same feature will cause the accuracy of certain algorithm to raise while another algorithm to down. But in general, RF algorithm seems to perform better, reaching the highest accuracy at 62% which is comparable to some complex machine learning algorithms. The other two algorithms reached around 50% accuracy at this dataset, which doesn't work really well as expected.

We also tried some other algorithms like Passive Aggressive (PA), the classical online algorithm for multi-class classification. But it seems to be under fitting, yielding only one label output, which leads to a low accuracy. Maybe some methods like normalization or scaling on data are needed to implement this algorithm.

VI. FUTURE WORK

Though getting relatively good results using only plain machine learning methods, there are still lots of things for us to develop. Various features including player statistics, manager index, match weather and even the mood of teams can be added to perform a more complex system which may get better prediction accuracy. Additional features will take hard work to collect, so the data source and corresponding crawler is another important part. Besides new features, some advanced machine learning algorithms may produce relatively higher accuracy. The most challenging thing may be that they are hard to be scalable, like Neural Network or Genetic programming. These algorithms are proved working well in the predictive model, but they heavily depend on iteration and are hard to become scalable.

Because our research is based on data mining, it can be easily expanded to other applications in various areas. The pipeline we built didn't rely on the data format, and can be applied to big data as it exploited Hive as its underlying. Then given appropriate features, our prediction model can be used for basketball, baseball or even non-sports prediction. What's more, we can include auto feature selection algorithm to make it more flexible to be applied to other areas.

VII. CONCLUSION

In this paper, we have proposed a predictive system for football match result prediction. This system contains a full pipeline from data collection to result prediction. All the data processing and machine learning parts are based on Apache Hive, so it supports large amount of data.

As experiments, we have tried several combinations of features, together with several well-known multi-class classification algorithms. Results have shown that feature selection is essential for machine learning, and not all of the algorithms will fit the data we have. Though the accuracy may not be really high, the performance is still satisfying in such a scalable platform. Besides this, some interesting insights can be drawn from this investigation. First of all, what really dominates the match result is the goal difference in the history and the times they won as the home/away team. And what's surprising is that the possession rate of a team seems important too. Secondly, adding too much features isn't always a good thing, it may even get things worse and lower the accuracy dramatically. So the feature selection is really important in such a predictive system. Finally, a complex system like a football match is hard to predict. Though many researchers and companies paid lots of time exploring the new algorithm, it seems that they haven't got a satisfying result over a large dataset. That's still encouraging, which means there remains a fantastic world for us to explore.

REFERENCES

[1] T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

[2] Apache Hivemall. https://hivemall.incubator.apache.org/.

[3] I. C. Peace, N. E. Okechukwu. An Improved Prediction System for Football a Match Result. IOSR Journal of Engineering, Vol. 04, Issue 12, 2014.

[4] D. Buursma. Predicting sports events from past results: Towards effective betting on football matches. *Preceding 14th Twente Student Conference on IT. University of Twente, Faculty Electrical Engineering, Mathematics and Computer Science, Netherlands*, 2011.

[5] C. K. Leung, K. W. Joseph. Sports data mining: predicting results for the college football games. *18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems* - KES 2014.

[6] F. Louzada, A. K. Suzuki, L. E. B. Salasar. Predicting Match Outcomes in the English Premier League: Which Will Be the Final Rank? Journal of Data Science 12(2014), pp 235-254.

[7] A. Gangal, A. Talnikar, A. Dalvi, V. Zope, A. Kulkarni. Analysis and Prediction of Football Statistics using Data Mining Techniques. International Journal of Computer Applications. Vol. 132 - No. 5, December 2015.

[8] K. Crammer, A. Kulesza, M. Dredze. Adaptive regularization of weight vectors. *Advances in Neural Information Processing Systems, 22:414–422, 2009.*

[9] B, Leo. Random Forests. *Machine Learning.* **45** (1): 5–32.

[10] J. Wang, P. Zhao, S. C. H. Hoi. Exact Soft Confidence-Weighted Learning. *in Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 121–128.*