International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 4321~4325
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp4321-4325

Comparing machine learning and ensemble learning in the field of football

Shuaib Khan, Kirubanand V.B
Department of Computer Science, CHRIST (Deemed to be University), India

Article Info

Article history:

Received Des 7, 2018
Revised Apr 13, 2019
Accepted Apr 25, 2019

Keywords:

Decision trees
Ensemble learning
Prescision
Support vector machines
XGBoost

ABSTRACT

Football has been one of the most popular and loved sports since its birth on November 6th, 1869. The main reason for this is because it is highly unpredictable in nature. Predicting football matches results seems like the perfect problem for machine learning models. But there are various caveats such as picking the right features from an enormous number of available features. There have been many models which have been applied to various football-related datasets. This paper aims to compare Support Vector Machines a machine learning model and XGBoost an Ensemble learning model and how Ensemble Learning can greatly improve the accuracy of the predictions.

Corresponding Author:

Shuaib Khan,
Department of Computer Science,
CHRIST (Deemed to be University),
Hosur Main Road, Bhavani Nagar, S.G. Palya, Bengaluru, Karnataka 560029, India.
Email: shuaib.khan@cs.christuniversity.in

1. INTRODUCTION

Football is a very unpredictable sport, the number of upsets caused by weaker teams beating relatively stronger teams is boundless, maybe why the sport is loved by so many all across the world. When it comes to who's going to win in a football match, there is a whole industry around it, pre-match analysis by football pundits and experts, post-game analysis by former players or professionals, Entire channels like ESPN , Sky Sports are dedicated to trying to analyse and figure out as to which team is going to win the match and even during halftime, there are commentators trying to predict who is going to win based on half-time stats. Betting Companies thrive on the unpredictability of football matches. There are various betting companies who have their own ways or models to predict the results of these matches, based on the prediction of these models they can adjust their odds accordingly.

There have been many papers and models implemented to predict the matches, most of which have achieved a reasonable amount of accuracy. The objective of this paper is to show the difference between a Machine learning model and an ensemble learning model. Machine Learning can be applied to the various aspects of real-life. But every application of a Machine Learning is different as there is a vast variety of data generated in the modern day. For example, a machine learning model used to predict the value of bitcoin might not be very accurate in classifying pictures of dogs. Choosing the right machine learning model is a part of the problem. The other part is getting the data prepared for the model. Often times we do not get the ideal data set, there may be missing values, duplication, etc. Pre-Processing the data set is the other part of the problem.

The performance of Machine Learning models and classifiers are usually ranked on some form of Accuracy. That is the comparison between the actual results and the predicted or obtained results. Ensemble learning aims to improve the accuracy of your learners (classifiers) by assembling them together. The errors

produced by a Machine Learning classifier can be broken down into bias, variance and Irreducible error. Ensemble learners help us get the right balance between bias and variance errors. This balance is also known as the Bias-Variance trade-off. In this paper, we look at one type of ensemble learning model called Boosting. Boosting is iterative in nature and adds weight to an observation or data based on the previous results of classification.

2. RESEARCH METHOD

2.1. Related work

Joseph and Fenton [1]: Bayesian Nets have been used in this paper to anticipate the results of Soccer matches and the result is compared with other models such as K-Nearest-Neighbours, Mc4 etc. The paper uses expert opinion for feature selection instead of mathematical models and the analysis is done on the matches played by Tottenham Hotspur. The paper shows that the Bayesian Nets outperform the other classifiers when the data set is disjoint.

Dobravec [2]: This paper recognizes the difficulty of the machine learning approach in the field of football. This paper uses a Matrix Factorization Model which forecasts the amount of goals scored by a team against a certain opponent. The AUC score obtained by the model is 0.677.

Dušan and Diana [3]: This paper uses Statistical Techniques such as Poisson distribution to predict football matches. The model applies Poisson distribution to the first half of the season and then using the results it simulates the other half of the season's results. Their Model can be used for a priori impact analysis by going through simulations of different management strategies based on their expected effects on match results.

Haghighat et al. [4]: This paper identifies two main problems for data mining in the field of Football. The first is the relatively low accuracy of classifiers trying to predict data, implying more accurate models need to be found and the second is the lack of good quality free data sets in terms of the statistics. Most of the datasets contain data collected from websites and not actually relevant statistics.

Forrest and Simmons [5]: This paper goes over the quantitative factors affecting the beautiful games, it tries to establish a relationship between the Home Team Supporters and the influence on the referee. It establishes that the home team, in general, receives fewer Yellow Cards.

Alejandro et al. [6]: This paper speaks about the Home advantage phenomenon teams face while playing at home, it tries to explain or find a reason for this phenomenon. It concludes by saying that it can be a combination of factors such as behaviour of the crowd, phycological effect of the players, familiarity with the stadium etc.

2.2. Methodology

2.2.1. Data cleaning and pre-processing

The Dataset selected contained features such as the number of goals scored by home team, the number of goals scored by away team, Shots taken by home team, Shots taken by away team, home team points, away team points, a variety of betting odds, and finally the Full-time result. The datasets collected were from the year 2000 to 2013.

Selecting the right features is a very important part of Machine Learning, these can be done using Statistical tests such as Pearson's Correlation, Linear Discriminant analysis, ANOVA, Chi-Square tests etc. However, in this model, we computed additional features from the dataset itself such as Home Team Win Streak, Home Team Loss Streak, Away Team Win Streak, Away Team Loss Streak, Difference in Points from the Home Team and the Away Team, Difference in Last Year's Predictions. All these features were the dependent variables. Full-Time Result was taken as the Independent Variable.

2.2.2. Support vector machines

Support Vector Machines has been considered as one of the go-to algorithms for data scientists, but why is it a favourite? This is because of one Reason, The Kernel Trick [7]. SVM is an Efficient Data Analysis Algorithm or Model which can be used both for classification as well as well as regression. It uses a hyperplane to separate the data into classes. This hyperplane or line must be selected in such a way that it maximizes the distance between the closest data point of each class. It is crucial that we find an optimal hyperplane because it classifies the data correctly and we will have higher accuracy on unseen (test) data. To find the optimal hyperplane we need a margin. A margin is the distance between the closest point and the hyperplane. The Margin is called a no man's land because there shouldn't be any data points between the hyperplane and the margin. As shown in Figure 1.

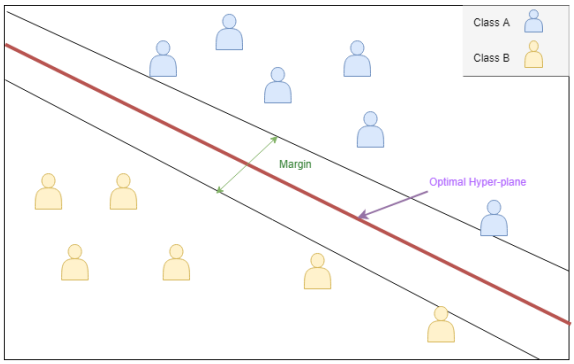

Figure 1. Depiction of SVM

The Kernel Trick is used in case the data is Non-Linear. The Kernel trick coverts our data (usually to a higher dimension) in such a way that we can draw build an optimal hyperplane. In other words, it converts the data into unrecognizable data which can be used by the SVM. This helps to accurately draw a margin between classes. A kernel function is responsible for transforming the data. There are many Kernels functions for almost all types of data. The Kernel function used here is called RBF (Radial Basis Function) as shown in Figure 2.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Figure 2. RBF formula

Here ||x-x'||2 is the Euclidian distance between two data points and σ is a parameter. The RBF kernel computes the distance from the origin or any other called a center. It is a real valued function which is used get an approximation of functions [8].

2.2.3. XGBoost

The Ensemble Learner used in this model was a boosting algorithm known as XGBoost. Boosting is a type of ensemble learner that trains the model on a randomized sample of the data and for the data points which weren't predicted correctly, it includes them in the next sample of randomly selected data. In other words, it adds weight to the unsuccessful predictions and trains the classifier again.

Boosting trains, the classifiers in a sequence in such a way that a new classifier should concentrate on those cases which were classified incorrectly. The results of the sequence of classifiers are compared and a voting to determine the output as shown in Figure 3.

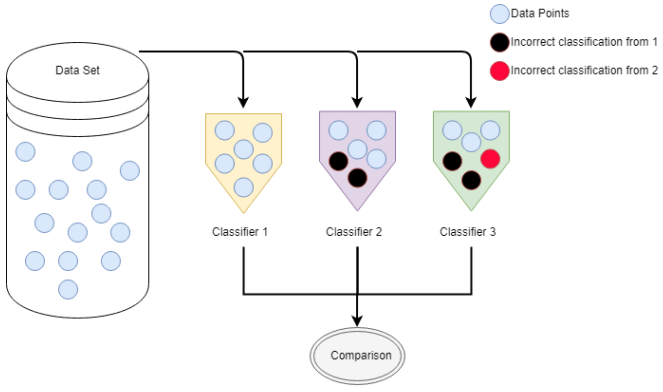

Figure 3. Depiction of boosting

XGBoost stands for extreme gradient boosting. Gradient Boosting works on the principle of a type Decision trees known as Classification and Regression Trees or CART. Trees are good at handling huge datasets, they can handle qualitative as well as quantitative data, they can ignore redundant variables, but one major drawback is that the prediction performance is very poor, but this is because of a large amount of variance. XGBoost solves this problem by taking a specific number of trees, each tree is grown (trained) to the weighted versions of the training data. This form of weighting decorrelates the trees that is it removes the correlation between trees by focusing on the regions missed by the past trees. The final Classifier is the weighted average of the classifiers. Gradient boosting improves all the good features of trees such as variable selection, mixed predictors etc. and improves on the weak features such as prediction performance and scalability of trees [9].

3. RESULTS AND ANALYSIS

As mentioned before, the goal of this paper is to compare an ensemble learning model and a machine learning model to show how an ensemble learning can greatly improve accuracy. First, the data is cleaned and Features are computed and added using the existing data set. The features are then selected put into the SVM model with the RBF kernel. This data is also fed into the XGBoost model and the results from both the models are compared as shown in Figure 4.

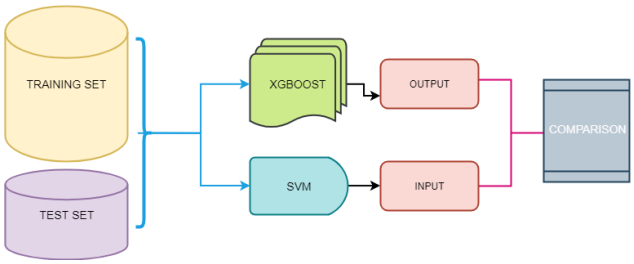

Figure 4. Model used for comparison

Both Accuracy score and F1 score are used to calculate the performance of the models. The F1 score uses precision and recall. It maintains the balance between the precision of the output and the recall of the output.

Let,
fp = false positives;
tp = true positives;
tn = true negatives;
fn = false negatives.
Precision = tp/ (tp + fp)
Recall = tp/ (tp + fn)
F1_score = 2 * (Precision*Recall)/(Precision+Recall)

The Observations obtained are as shown in Tabel 1 and Table 2:-

Table 1. SVM results

Variable	F1 score	Accuracy
Training Set	0.715	0.756
Test Set	0.654	0.660

Table 2. XGBoost results

Variable	F1 score	Accuracy
Training Set	0.855	0.869
Test Set	0.801	0.824

4. CONCLUSION

The Goal of this paper was to show the superiority of ensemble learning over machine learning models in the field of Football. As we can see, XGBoost performs significantly better than its machine learning counterpart SVM. With an accuracy score of 0.855 and an F1_score of 0.801. It should be noted that these predictions or classification were made with the use of in-game statistics which are not available before the match takes place. This is because the aim of this paper is not to predict the football matches or come up with an algorithm or model to predict the football matches. There are many different ensemble models and machine learning models which can be implemented in this are to predict football Matches. The Results of this paper shows that ensemble learning can be a good choice when trying to predict the results in this field.

ACKNOWLEDGEMENTS

The author would like to thank facilitators of the university have been a most helpful ally in structuring the data and understanding the problem domain. The paper would not have been complete without the massive contributions by all the faculties and resources in lieu of Christ (Deemed to be University), for both research and review.

REFERENCES

[1] A. Joseph, N. E. Fenton, "Predicting Football results using Bayesian Nets and other Machine Learning Techniques," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 544-553, Nov 2006.
[2] Stefan Dobravec, "Forecasting Football World Cup Results using a Matrix Factorization Technique," *Elektrotehniski Vestnik/Electrotechnical Review*, vol. 82, no. 1, pp. 61-65, Jan 2015.
[3] Munđar Dušan and Šimić Diana, "Croatian First Football League: Teams' performance in the championship," *Croatian Review of Economic, Business and Social Statistics*, vol. 2, no. 1, pp. 15-23, 2016.
[4] Maral Haghighat, Hamid Rastegari, Nasim Nourafza, "A Review of Data Mining techniques for Result Prediction in Sports," *ACSIJ*, vol. 2, no. 5, Nov 2013.
[5] Buraimo B., Forrest D. and Simmons R., "The 12th man?: refereeing bias in English and German soccer," vol. 173, pp. 431-449, Mar 2010.
[6] Legaz-Arrese, Alejandro, Moliner-Urdiales, Diego, Munguía-Izquierdo, Diego, "Home Advantage and Sports Performance: Evidence, Causes and Psychological Implications," *Universitas Psychologica*, vol. 12, no. 3, pp. 933-943, 2013.
[7] Ben Ulmer, Mathew Fernandez M peterson, "Predicting Soccer Match Results in the English Premier League," 2013, [Online]. Available: http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez, %20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf.
[8] Hui Cao, Takashi Naito, Yoshiki Ninomiya, "Approximate RBF Kernel SVM and Its Applications in Pedestrian Classification," *MLVMA'08*, Marseille, France, Oct 2008.
[9] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD*, Aug 2016, 785-794.
[10] "2000-2013 Primier League Dataset," Footba-data.co, (2018), [Online]. Available: http://www.football-data.co.uk/englandm.php

BIOGRAPHIES OF AUTHORS

Shuaib Khan, Corresponding Author, (Student) Department of Computer Science, Christ (Deemed to be University), Bangalore, India, Email: skpalardin@gmail.com.

Kirubanand V.B (Associate Professor) Department of Computer Science, Christ (Deemed to be University), Bangalore, India.