Football Match Winner Prediction

Saurabh Vaidya
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Harshal Sanghavi
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Kushal Gevaria
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

ABSTRACT
Prediction of football match outcome should follow approaches that are more generalized. Hence for our project we predict outcomes of English Premier League based on the historical data of the matches and using machine learning algorithms. We gathered data from past 10 seasons and extracted features like form, goals scored and conceded, shots ratio. The computation of form feature is different from has been prevalent till now. More focus is given to gain more insight and associate a deeper and better meaning to form of a team. Basic features like shots ratio and goals scored are combined to create feature of attacking quotient. We using Logistic Regression and implement voting algorithm between Random Forest and Naive Bayes classifier to achieve accuracy between 47-50% with mean absolute error of 0.37.

Keywords
Machine learning; Data mining; Prediction system; Football; Classifiers; Knowledge discovery database system

1. INTRODUCTION
2010 FIFA World cup, showed a display of sheer brilliance by Paul the Octopus. Paul predicted the winner correctly an astonishing 8 times when he was tested. There are other predicting techniques, which can predict the outcome after half-time; while some predict the outcomes on an on-going basis; however, the accuracy is not good. So, for the love of the game and the eagerness to learn new techniques of prediction, we have made an attempt to devise our own method to predict the outcome of a football match.

The problem of predicting football match winner is a multi-class classification problem having three classes: win, loss, draw. Out of these, win and loss are comparatively easy to classify. However, the class of draw is very difficult to predict even in real world scenario. A draw is not a favored outcome for pundits as well as betting enthusiasts.

English Premier League (EPL) is the most watched football league in the world with almost 4.7 billion viewers. In our paper, we have chosen English Premier League for its competitiveness as well as its random nature of outcomes. For example, in the season of 2010-11, the distribution of wins, losses and draws was 35.5%, 35.5% and 29% respectively. So if we calculate the measure of randomness:

$$\text{Entropy} = -(.29 * \log3(.29) + 2(.355 * \log3(.355)))$$

$$= 0.72 \text{ [3].}$$

This is very close to 1 (state of complete randomness). Thus testing our results on EPL would only help to justify the generality of our approach.

The major challenge in task of predicting match outcome is the extraction and availability of required data. The data source used by us in this project is www.football-co.uk . The data has to be scraped and stored to extract the features. We collect data over 10 seasons from 2004-05 to 2014-15. We extract set of 4 features per team. All the data are scraped with help of crawlers.

The features generally used are taken in its direct form like shots, cards, goals etc. However, we have attempted to perform some computations to make some complex features. Various machine learning techniques have been used to predict match outcomes like Clustering, SVM, Bayesian classifiers etc. We would be trying different techniques to find the one which suits our data sets.

2. LITERATURE REVIEW
The term "Data Mining" was first used around 1990 in the database community. Data mining and Knowledge discovery are used interchangeably. Data mining is the process of extracting information from a data set and converts it into understandable structured form [4]. Data mining has many applications and thus this term is much useful in predicting the match winner in football sports by analyzing the previous match data. Data mining with machine learning can make such predictions work efficiently. Arthur Samuel in 1959, defined machine learning as "Field of study that gives computers the ability to learn without being explicitly programmed". Machine learning conflated with data mining helps us to focus more towards exploratory data analysis. Based on trained data, machine learning does the prediction that depends on the properties learnt from those trained data [5].

Betting is widely popular among sporting events ranging from cricket, football to tennis and snooker. Douwe Buursma gives importance towards effective betting on football matches [1]. Betting is prominently popular in football, as it is one of the world's famous and most widely watched sport in the world. The betting system works in following way: The bettor wins money if his bets placed turn out to be correct and loses money otherwise. The money earned or lost is based on the odds determined by the bookmakers. When the probability of the outcome is say 0.5, the bookmakers odds would be 5. However to earn profit, the bookmakers place the odds at say 4.5. Thus, to eliminate this "unfairness" it is necessary to find accurate probabilities of wins or draws to beat the bookmakers' odds. Douwe Buursma uses different machine learning classifiers and the accuracy of 55.08% is obtained by using regression and multi-class classifier [1].

Nivard van Wijk uses the betting concept which leads one to predict a match winner and thus proposes two models to explain the prediction. These two models are toto-model and score-model respectively. This paper explains the prediction system mathematically by all the methods and formulas specified in the article itself. The accuracy of about 53.03% is

obtained after comparing all the models proposed in this paper [2].

Ben Ulmer and Matthew Fernandez predicted the soccer match results in English Premier League. They used some machine learning techniques, which include classifiers namely Linear from stochastic gradient descent, Naïve Bayes, hidden Markov model, Support Vector Machine and Random forest. Accuracy of each and every model was calculated to find the better approach. They proposed that the results of the first few matches couldn't be predicted due to the lack of data regarding the form of the team. They compared all the methods out of which SVM showed the best result of 40% - 52% accuracy [3].

3. WORKING OF THE SYSTEM

As seen in literature survey, different systems had their own different set of parameters and classifiers. The accuracy of the system would thus depend on the feature selection and computation as well as the type of classifier used. In order to achieve a better accuracy than previous systems, we would focus on selecting proper features and computing accurate algorithms on those features and selecting the best classifier. The prediction system proposed by us would have three main parameter components viz. current form, attacking quotient and defensive quotient.

The current form is calculated keeping in mind two factors: home/away outcome and relative position of two teams. A form matrix is constructed which implements the above factors and gives a detailed information about the magnitude of a team's loss or win.

Table 1. FORM MATRIX

Teams	Points	Multiplying Factor	Home loss	Away win
A	0.75	0.15	-20%	20%
B	0.6	0.25	-16%	16%
C	0.4	0.4	-12%	12%
D	0.15	0.6	-10%	10%

The above table is used to calculate a team's form (recent 5 matches). 20 teams are divided equally in groups of 4 based on their table position. When a team wins, +1 and some extra points are awarded which depicts the magnitude of that win. That magnitude is calculated using the above table. For example, if a team from group A wins against a team of group C (home of group C), points structure of Team A will be

Points = ((+1) + (0.15 * 0.4)) * 1.2

And that of team C will be Points = ((-1) – (0.15 * 0.4)) * 0.88

Finally, all the points of 5 recent matches will be added to generate a collective form.

Two main aspects of a football game are attack and defense. Thus comparing these two quotients of two teams gives us an intuition about the better team both attack-wise and defense-wise. The attacking quotient is again computed using following features: shots on target and shots on target/goals ratio. These two features would signify how good the team is in terms of attack. The defense quotient is computed using the features: successful tackles and intercepted passes. These would signify the strength of the defense.

After feature selection and computation, the next task would be selecting upon the classifier to be used. Initially we used Logistic regression to classify the data set, however it classified only 2 classes and not the 3rd one.

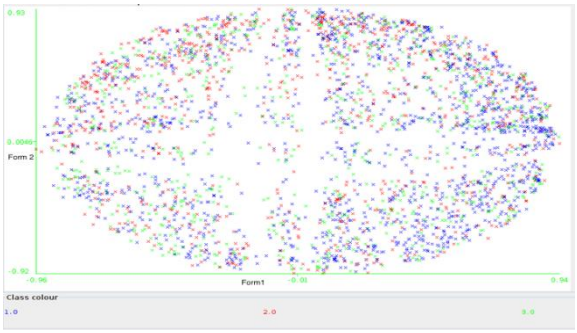

Fig. 1: Form v/s Form Graph

On plotting the dataset on a graph, we got the following result:

As we can observe, the dataset is very sparse and hence using. Decision trees and Naïve Bayes classification would yield better results. Hence, the next algorithm that we implemented is Vote algorithm. This algorithm uses the best outcomes of all the listed algorithms and generates a cumulative outcome. We used Random forest and Naïve Bayes classification algorithms. This algorithm was able to classify the 3rd class which was not possible using any other algorithm.

The following is our system architecture:

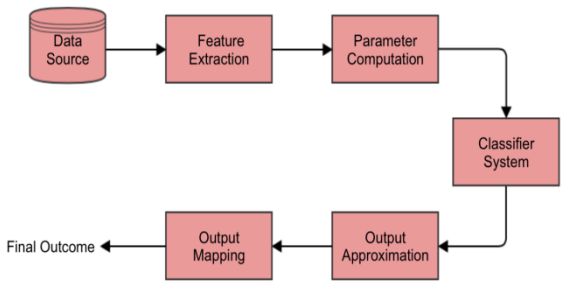

Fig. 2: System Architecture

As seen in the architecture we would extract all our features that would be required, from a data source and compute the above-mentioned parameters such as form and attack, defense quotients. The classifier system would give us a value that will determine the class to which the output would belong. This output would then be approximated and mapped to defined outputs (1 for win, 0 for a loss, and 0.5 for a draw). The final output would be a list of outcomes predicted for a set of matches.

4. EXPECTED OUTCOME

We collected data from various websites and data sources using different scrapping tools. We generated a mathematical model to represent the data in the format required by the algorithms. The dataset was then divided in the ratio 80:20 (training: testing). We achieved 49.37% accuracy using Logistic regression algorithm and below is the confusion matrix:

Table 2. Confusion Matrix of Logistic Regression

	Predicted Win	**Predicted Loss**	**Predicted Draw**
Actual Win	268	32	1
Actual Loss	135	57	0
Actual Draw	138	27	0

As we can see from the confusion matrix, Logistic regression classifies only 2 classes and just 1 instance of class 3. Hence, we used a different algorithm Vote which selects the best results of multiple algorithms. Here, we have used Random forest and Naïve Bayes classification algorithms for voting. Accuracy achieved is 47.11% and below is the confusion matrix:

Table 3. Confusion Matrix of Vote Algorithm

	Predicted Win	**Predicted Loss**	**Predicted Draw**
Actual Win	235	52	14
Actual Loss	114	66	12
Actual Draw	112	44	9

Although this algorithm is not as accurate as the previous one, it still classifies the 3rd class and hence there is a compromise between accuracy and classification of all classes.

5. CONCLUSION AND FUTURE SCOPE

Thus, it is seen that the case of draw reduces the accuracy of predicting the remaining two classes. It is observed that by removing the draw instances, accuracy can be increased up to 65%. Logistic regression fails to classify the draw class. So in order to achieve generality, voting algorithm is preferred. Availability of more features that can help in solving the issue of predicting draw class would improve the accuracy. Also, algorithms optimal for sparse data such as decision trees and boosting algorithms may also increase the accuracy.

6. REFERENCES

[1] Douwe Buursma; Predicting sports events from past results, University of Twente, 2011.

[2] Nivard, W. & Mei, R. D.Soccer analytics: Predicting the of soccer matches. (Master thesis: UV University of Amsterdam), 2012.

[3] Ben Ulmer and Matthew Fernandez; Predicting Soccer Match results in the English Premier League, cs229, 2014.

[4] Data mining [Online]. Available: https://en.wikipedia.org/wiki/Data_mining

[5] Machine Learning [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning