International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 10, October 2015)

# Football Match Winner Prediction

Kushal Gevaria[1], Harshal Sanghavi[2], Saurabh Vaidya[3], Prof. Khushali Deulkar[4]
*Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*

***Abstract*— A common discussion subject for the male part of the population in particular, is the prediction of next weekend's soccer matches, especially for the local team. The football data available from different sources and the use of it has become popular day by day. The problem of modeling football data has become increasingly popular in the last few years and many different models have been proposed with the aim of estimating the characteristics that bring a team to lose or win a game, or to predict the score of a particular match. Knowledge of offensive and defensive skills is valuable in the decision process before deciding the outcome of the match. In this study, the methods of machine learning and data mining are used in order to predict outcome of football matches by determining the winning team based on data available from previous matches. Although it is difficult to take into account all features that influence the results of the matches, an attempt to find the most significant features is made, also few software programs are used to find out the features that are most contributing such as WEKA. Various classifiers, like logistic regressions, SVM, Bayesian networks etc. used, are tested to solve the problem. Finally the most significant features are proposed and the way of computing new parameters from these features is proposed and the best classification method is to predict the outcome.**

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

## I. INTRODUCTION

In 2010 World cup, there was a display of sheer brilliance by Paul the Octopus. The sea dweller predicted the winning team correctly 8 times when he was tested. There are other predicting techniques, which can predict the outcome only after half time; 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.

Prediction system usually works by learning from the past for which the data is gathered. The data available follows a particular pattern. The job of a prediction system is to observe the data and determine the pattern so as to predict the future results. The prediction system is not as easy as it seems. Complexity is its intrinsic characteristic. It uses various Artificial intelligence methods and techniques for better results.

One of the applications of prediction system is in stock markets where it can be used to predict the rate of various stocks.

The efficiency of a prediction system can be determined with the help of accuracy. Accuracy is basically probability or likelihood of occurring of an event. More the probability better is the accuracy. The advancements in technology have helped the prediction system to achieve accuracy in the range of 75%-80%. Such a high probability of the occurrence of an event shows the great heights to which humans have reached in terms of technology.

Prediction of the result of a football match is another upcoming application. The system predicts whether a team will win or lose a particular match. This prediction can be done using machine learning. Various classifiers (algorithms or methods) that can be used are Linear SVM, Logistic regression, Random forest, stochastic gradient descent, Hidden Markov model etc. [1]. The accuracy of the system varies from classifier to classifier and features to features. Features are basically the parameters that affect the outcome of a football match. Form of a team in the last few matches is one of the important features. If a team has been losing all the 5 matches, there is a high possibility that it will lose the next match and same is the case of winning. However, form is not the only feature to be considered. Various others are: shots on target, goals conceded, red cards, yellow cards, injury of main players, home and away matches, goal difference etc. It is also observed that the data available in games such as FIFA are very close to the actual ones. These attributes include agility, free kick, corner kick, dribbling and many such features of individual players. Each player is assigned points for each of the attributes, which can be used as a parameter to compare the efficiency of the players at different point in the game [4]. This comparison can be used as one of the features for prediction. Combination of various features results in varying accuracies. The best combination can be selected on the basis of these accuracies.

## II. LITERATURE REVIEW

The term "Data Mining" was first used around 1990 in the database community. Data mining and Knowledge discovery are used interchangeably.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 10, October 2015)

Data mining is the process of extracting information from a data set and converts it into understandable structured form [7]. 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 [8].

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].

Albina Yezus used data set from two sources to predict the football match outcome. The objective of this paper was to achieve maximum accuracy. Classifiers used were Random forest and K nearest neighbor.

Accuracy obtained from these two models was 63.4% and 55.8% respectively. Albina suggested that in order to achieve high accuracy, practical implementations like SVM and Linear Regression could be used [3].

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 69% - 55% accuracy [4].

Focusing more towards complex data set, Igiri Nwachukwu predicts the football match winner with the help of a data mining tool namely rapid miner as well as including more and more features possible. Knowledge Discovery in Database (KDD) helps in gathering as many features as possible. Classifiers used are artificial neural network and logistic regression. The accuracy of about 93% is obtained in predicting the match winner in this article [5].

Jongho Shin and Robert Gasparyan (2015) predicted the match result using data from virtual games like FIFA. They made the comparison between their predicted output and the data produced by predictors using real time data. Various attributes of the individual players were combined and compared that with the real time prediction. Accuracy obtained using real time predictor is 75% and using virtual predictor is 80% [6].

## III. Comparative Study

Given below is the tabular representation of comparative analysis on different types of techniques. In review of the literature, we learnt about various techniques used for prediction of the winner. These techniques involves combination of different parameters and appropriate classifiers which gives varying accuracies. We have presented the comparative study of these techniques in the table below:

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 10, October 2015)

**Table I**
**Comparative Study Of Different Types Of Techniques**

| Author | Classifiers | Parameters | Accuracy |
|---|---|---|---|
| Douwe Buursma [1] | Classification via regression | Goals scored, Goals conceded, Average points per match, Number of wins (home and away) | After testing of all classifiers, the best average accuracy obtained was 55.08% using Classification via regression and multi-class classifier |
| | Multi-class classifier | | |
| | Rotation forest | | |
| | Logit boot | | |
| | BayesNet | | |
| | Naïve Bayes | | |
| | Home wins | | |
| Nivard van Wijk [2] | Random probability and team grouping | Frequency of number of goals per match, Average number of points per match | 48% |
| | Multi independent model | | 47.24% |
| | Single independent model | | 53.03% |
| | Dependent model | | 53.55% |
| Albina Yezus [3] | Random forest | Result, goal difference, table position, history of results, form, motivation, concentration | 63.4% |
| | K nearest neighbor | | 55.8% |
| Ben Ulmer [4] | Baseline | Results of the matches of past 10 years | 60%-40% |
| | Naïve base | | 48%-44% |
| | Hidden markov model | | 48%-44% |
| | SVM | | 69%-55% |
| | Random forest | | 51%-49% |
| | Stochastic gradient descent | | 51% |
| Igiri Nwachukwu [5] | Artificial neural network | Home advantage, injury of main players, external cup influence, Home and away goals, Home and away shots, Home and away corners etc. | 85% |
| | Logistic regression | | 93% |
| Jongho Shin [6] | Logistic regression | Shots on target, goals scored, red cards, yellow cards, 33 attributes of each player from FIFA | 75% |
| | Linear SVM | | |
| | RBF SVM | | 80% |
| | SGD | | |

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 10, October 2015)

## IV. Proposed 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. Thus a team higher up the table winning against a lower team and higher team winning against another top table team always mean different. Similarly a top team losing at home against another top team and top team losing at home against a bottom team would mean different. Thus a current form is simply not a measure of outcomes but also how valuable or meaningful those outcomes were.

Two main aspects of a football game are attack and defence. Thus comparing these two quotients of two teams gives us an intuition about the better team both attack-wise and defence-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 defence quotient is computed using the features: successful tackles and intercepted passes. These would signify the strength of the defence.

After feature selection and computation, the next task would be selecting upon the classifier to be used. As seen in the comparative study, different classifiers resulted in different accuracies. Also no system relied upon a single classifier. Thus we would also test our features with different classifiers such as Support vector machines, Bayesian Networks, Logistic Regression etc.

*The following is our system architecture:*

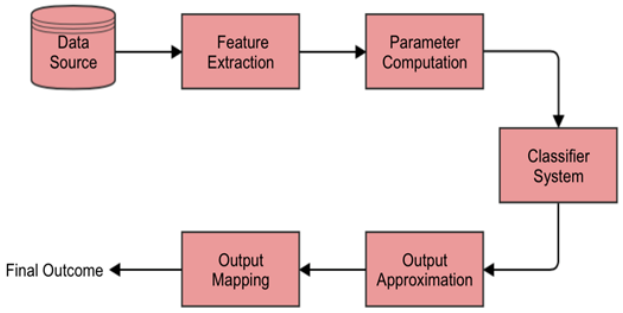


*Figure 1 System Architecture*

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.

## V. Conclusion

A Comparative study helps to determine which technique turns out to be the best of all. Classifier named logistic regression predicts the outcome of a football match with the precision of more than 90%. Also on the same set of features Artificial Neural Networks gave 85% accuracy. Rest other systems range from 47%-85% accurate. Thus it can be concluded that the accuracy of a system would depend on selecting right features, computing values of parameters accurately and selecting best classifier algorithms. The challenging task in football match prediction is the accuracy with which we can predict correct output and by advancements in data mining and machine learning we would get further improved results.

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] Albina Yezus; Predicting outcomes of Soccer matches using machine learning, Saint-Petersburg University, 2014.

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

[5] Igiri, Chinwe Peace. Nwachukwu, Enoch Okechukwu; An improved prediction system for Football match result, IOSR Journal of Engineering volume:04, 2014.

[6] Jongho Shin and Robert Gasparyan; A novel way of Soccer match prediction, 2014.

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

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