International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 03 - Issue 07 || July 2018 || PP. 31-38

Analysis of Football Events

Pankaj Lathar
CBP Government Engineering College,
Jaffarpur, New Delhi - 110073

Abstract: An important feature of a sport competition is the uncertainty about the outcome. Competitive balance measures this degree of uncertainty about the result of a competition. High competitiveness means that there is high uncertainty about the rankings of teams or players. Classical measures of competitiveness are based on the ratio of wins of each team or other related measures. Comparison of ranking helps us in analyzing dynamic behavior of a sports league. There can be prediction of winning of a team, given the data about the opposite team, place, country etc. The winning is categorized in home team and away team. Predicting the winning, gives sponsors' and players, an idea about their winning and they can devise their strategy and practice accordingly. The rankings of the teams or players gives the competitiveness among them in the series. One can make use of the statistical method, known as Kendall's coefficient to analyze rankings. Kendall's coefficient gives us a global idea about the relative number of times any two possible competitors have exchanged their respective positions through the corresponding family of rankings. The value of Kendall's coefficient determines the degree of competition between two series of rankings..
Keywords: European football (soccer) leagues, Naive Bayes classifier, C5.0 ,Kendall's coefficient.

1. INTRODUCTION

An important feature of a sport competition is the uncertainty about the outcome. Competitive balance measures this degree of uncertainty about the result of a competition. A high competitiveness means that there is high uncertainty about the teams ranking. Classical measures of competitiveness are based on the ratio of wins of each team or other related measures. Comparison of ranking helps us in analyzing dynamic behavior of a sports league.

Most publicly available football (soccer) statistics are limited to aggregated data such as Goals, Shots, Fouls and Cards. When assessing performance or building predictive models, this simple aggregation, without any context, can be misleading.

A football game generates much more events and it is very important and interesting to take into account the context in which those events were generated. This dataset should keep sports analytics enthusiasts awake for long hours as the number of questions that can be asked is huge.

Aggregation data are the data combined from several measurements. In this case, the aggregated data is goal (is goal or not), events (red card, substitution etc.), shot outcome (on target, off target etc.) and many more. Traditional approaches include subjective prediction, objective prediction, and simple statistical methods. However, these approaches may not be too reliable in many situations. Data mining approach makes predictions based on a combination of four different measures on the historical results of the games.

2. DESIGN

The 5 European football (soccer) leagues dataset used in this research paper has been taken from the Kaggle[1]. The dataset provides a granular view of 9,074 games, totalling 941,009 events from the biggest 5 European football (soccer) leagues: England, Spain, Germany, Italy, and France from 2011/2012 season to 2016/2017 season as of 25.01.2017. Overall, over 90% of the played games during these seasons have event data. The dataset is organized in 3 files:

1) **Events.csv** contains event data about each game. Text commentary was scraped from: bbc.com, espn.com and onefootball.com

2) **ginf.csv** - contains metadata and market odds about each game. odds were collected from oddsportal.com

3) **dictionary.txt** contains a dictionary with the textual description of each categorical variable coded with integers[2].

The goal of our project is to build a predictive model for Foot ball events which helps to predict the winning team based on the predictions using previous years datasets. It helps to predict the goal by using the parameters such as player, team and event type, body part, location etc. It also helps to give ranking to each player and team based on their previous performances. We are using three modules namely:

1) Predictive module: Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modeling can be applied to any type of unknown event, regardless of when it occurred. In our Project we are applying it to the soccer game events. So we can predict the winning team in the upcoming games by using previous history of data. It has multiple sub modules like prediction of winning team by using naive bayes classifier , prediction of goals by C5.0 Algorithm etc.
2) Time series module **:**In time series module, we perform time based analysis for various events occurring in the game. In our project we have a dataset having multiple tuples for each games, each tuple representing an event. One can plot graphs depicting probability of occurrence of events at different time.
3) Ranking module: We are Performing Ranking module using Kendall's co efficient. It helps to give rank to each Team based on their performance.
The software used for the classification of this dataset is R. R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years [3]. The package used for Graphical user interface is Shiny in R. Shiny is an R package that makes it easy to build interactive web applications (apps) straight from R[4]. The software tool used for this predictive model is R- Studio. R-Studio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.
R-Studio is the premier integrated development environment for R. It is available in open source and commercial editions on the desktop (Windows, Mac, and Linux) and from a web browser to a Linux server running R-Studio Server or R-Studio Server Pro[3].The initial training set is 7500 and 2500 instances are used as the test data. We can do it by changing size of the train data and test data. For the prediction purpose two algorithms are used , Naive bayes classifier[5] predicts the winning of the team based on the parameters such as country, place, home team and away team and C5.0 algorithm[6] predicts the occurrence of goal based on the parameters having player, team, place, body part, shot location etc. For Ranking module we are using kendall"s coefficient[7].

3. IMPLEMENTATION

The traning data set is 7500 instances that is stored into the 'traindata' object and 2500 instances of test data set stored into the 'testdata' object. Based on the train data , test data is predicted. Now the algorithms are used for prediction and ranking.

3.1 C5.0

The classifier is tested first to classify unseen data and for this purpose resulting decision tree is used.C4.5 algorithm follows the rules of ID3 algorithm. Similarly C5 algorithm follows the rules of algorithm of C4.5. C5 algorithm has many features like-The large decision tree can be viewing as a set of rules which is easy to understand. C5 algorithm gives the acknowledge on noise and missing data. Problem of over fitting and error pruning is solved by the C5 algorithm and in classification technique the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification
The 'classification_model' object holds the classified data using the C5.0 algorithm initially for 8000 instances of the 'traindata' and the remaining 2000 is predicted. A graph is plotted for the actual and predicted accuracies. Repeat the same procedure to improve accuracy. Again a graph is plotted for the actual and predicted values and this time the errors are reduced to just 3.The accuracy obtained is 99% .

3.2 Kendall's coefficient

The Kendall rank correlation co-efficient evaluates the degree of similarity between two sets of ranks given to a same set of objects. This co-efficient depends upon the number of inversions of pairs of objects which would be needed to transform one rank order into the other. In order to do so, each rank order is represented by the set of all pairs of objects (e.g., [a,b] and [b,a] are the two pairs representing the objects a and b), and a value of 1 or 0 is assigned to this pair when its order corresponds or does not correspond to the way these two objects were ordered.

It is a co-efficient that represents the degree of concordance between two columns of ranked data. The greater the number of inversions the smaller the co efficient will be. Its range is from -1.0 to +1.0.
Kendall’s Tau is a non-parametric measure of relationships between columns of ranked data. The Tau correlation coefficient returns a value of 0 to 1, where:

- 0 is no relationship,

- 1 is a perfect relationship.

A quirk of this test is that it can also produce negative values (i.e. from -1 to 0). Unlike a linear graph, a negative relationship doesn't mean much with ranked columns (other than you perhaps switched the columns around), so just remove the negative sign when you're interpreting Tau[8].

3.3 Naive Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers

An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

4. TESTING AND COMPARISON

This research paper uses three algorithms. As seen in the implementation the C5.0 algorithm gives an accuracy estimate of 85-90% and Naive Bayes classifier algorithms where the accuracy is stuck around only 90%.

Higher the accuracy, greater is the efficiency of prediction of the winning team and goal. Initially the accuracy matrix is below 75 % after repeating the procedure it increases to 85-90% for all three algorithms :

Naive Bayes :
Country: Italy
Home team: Bologona
Away team: Hellas Verona
Prediction of winning: Bologona

Actual	Predicted	Freq
Away	Away	109
Home	Away	53
Away	Home	88
Home	Home	289

C5.0 :
Event type :Key pass
Shot Place : Bit Too high
Shot Outcome: On target
Locations : Attacking Half
Body Part: right foot
Situation :Open play
Fast break: 0
Prediction of occurrence of goal : No goal

Kendall's co efficient :

1) Ranking to players

Fig.1 Example of Ranking to player

players	X2012	X2013	X2014	X2015	X2016	X2017
gonzalo higuain	1	2997	2982	2969	257	3
rodrigo palacio	2	17	3039	77	681	1337
edinson cavani	3	1	14	8	248	1
marco reus	4	2982	2968	145	150	672
eden hazard	5	911	39	21	193	33

Fig.2 Accuracy matrix for Ranking to player

2) Ranking to teams

Fig.3 Example of Ranking to teams

teams	X2012	X2013	X2014	X2015	X2016	X2017
Real Madrid	1	5	4	1	4	8
Barcelona	2	1	6	2	2	15
Manchester City	3	8	5	7	13	11
Manchester Utd	4	3	21	16	16	21
Borussia Dortmund	5	16	14	43	8	30

Fig.4 Accuracy matrix for ranking to teams

Higher the accuracy, greater is the efficiency of prediction of the winning team and goal. After repeating the procedure accuracyt increases to 85-90% for all three algorithms.

5. RESULTS

PREDICTION OF WINNING TEAM

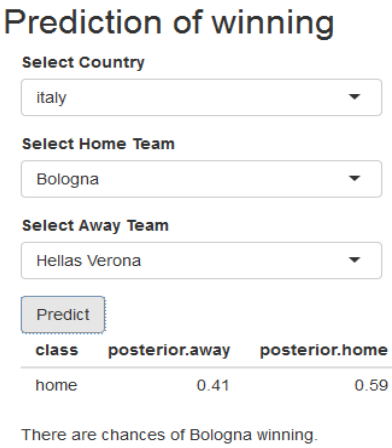


Fig. 5 Selection of parameters for prediction of winning

Actual	Predicted	Freq
away	away	109
home	away	53
away	home	88
home	home	289

Fig. 6 Accuracy matrix for prediction of matrix

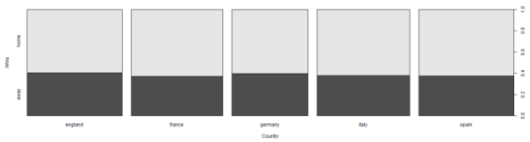


Fig. 7 Graph for prediction of winning model

PREDICTION OF OCCURRENCE OF GOAL

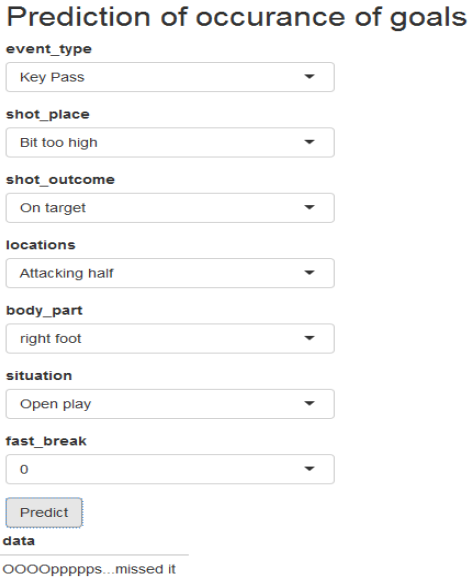


Fig. 8 Selection for prediction of occurrence of goals

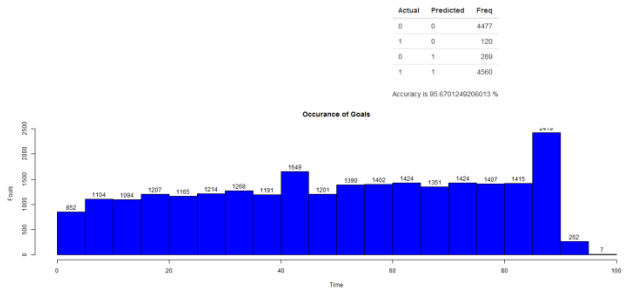


Fig. 9 Graph for prediction of occurrence of goals

PLAYER RANKING MODEL

Fig. 10 Selection of parameters for Player Ranking model

players	X2012	X2013	X2014	X2015	X2016	X2017
gonzalo higuain	1	2997	2982	2969	257	3
rodrigo palacio	2	17	3039	77	681	1337
edinson cavani	3	1	14	8	248	1
marco reus	4	2982	2968	145	150	672
eden hazard	5	911	39	21	193	33

Fig.11 Matrix for Player Ranking model

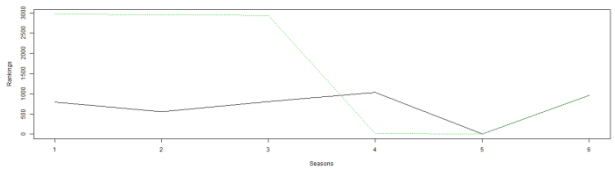


Fig. 12 Graph for Player Ranking model

TEMA RANKING MODEL

Fig. 13 Selection of parameters for Team ranking model

teams	X2012	X2013	X2014	X2015	X2016	X2017
Real Madrid	1	5	4	1	4	8
Barcelona	2	1	6	2	2	15
Manchester City	3	8	5	7	13	11
Manchester Utd	4	3	21	16	16	21
Borussia Dortmund	5	16	14	43	8	30

Fig. 14 Matrix for Team ranking model

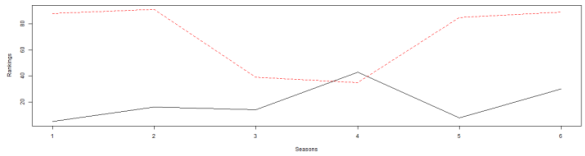
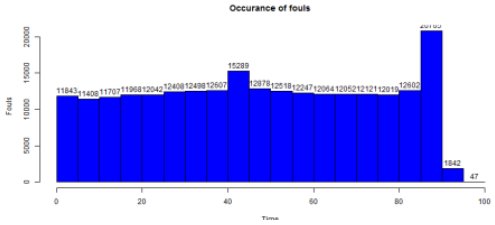


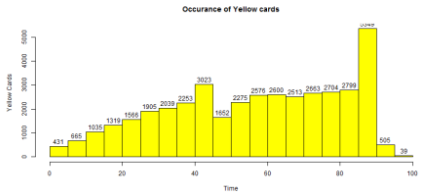
Fig. 15 Graph for Team Ranking model

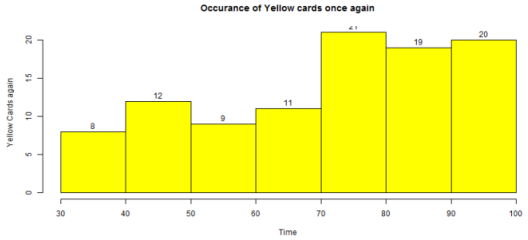
TIME SERIES MODEL

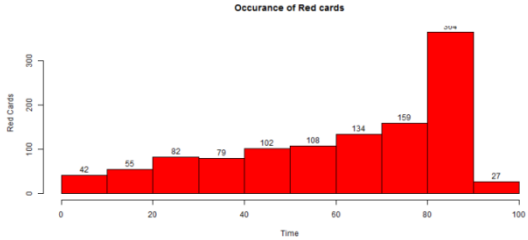
EUROPEAN FOOTBALL STATISTICS

An important feature of a sport competition is the uncertainty about the outcome. Competitive balance measures this degree of uncertainty about the result of a competition. High competitiveness means that there is high uncertainty about the rankings of teams or players. Classical measures of competitiveness are based on the ratio of wins of each team or other related measures. Comparison of ranking helps us in analyzing dynamic behavior of a sports league. There can be prediction of winning of a team, given the data about the opposite team, place, country etc. The winning is categorized in home team and away team. Predicting the winning, gives sponsors' and players, an idea about their winning and they can devise their strategy and practice accordingly. The rankings of the teams or players gives the competitiveness among them in the series. One can make use of the statistical method, known as Kendall's coefficient to analyze rankings. Kendall's coefficient gives us a global idea about the relative number of times any two possible competitors have exchanged their respective positions through the corresponding family of rankings. The value of Kendall's coefficient determines the degree of competition between two series of rankings. /n/n Below is the analysis about occurances of various events as per the data between year 2012-2016









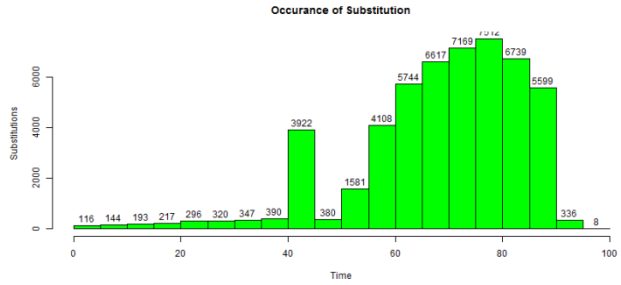


Fig. 16 Example graph for Time series model

6. SCOPE AND FUTURE WORK

An important feature of a sport competition is the uncertainty about the outcome. The current scope of the project aims at classifying and predicting based on the previous data to produce a effective predictive model which helps user to predict the winning team, occurrence of goal and ranking model which also helps to compare the efficiency between the two teams and also between the players based on the previous performance.

The development of this predictive model helps in the following situations in the future:
1) There can be prediction of winning of a team, given the data about the opposite team, place, country etc.
2) The winning can be categorized in home team and away team.
3) Predicting the winning, gives sponsors' and players, an idea about their winning and they can devise their strategy and practice accordingly.
4) The rankings of the teams or players gives the competitiveness among them in the series.
5) Kendall's coefficient gives us a global idea about the relative number of times any two possible competitors have exchanged their respective positions through the corresponding family of rankings.
6)Prediction helps sponsors to bid players by knowing their performance in home and away.
7) Based on the prediction, they come to know their week points so that they can improve their performance by working more on their week areas.
8) Prediction helps sponsors' to know which player has high performance rate in home and away team so that they can make a team of players having high performance rate in home when they need to play as home team and vice versa.
9) Based on the prediction sponsor or coach can know which player had more injuries in which season so that they can help him to overcome it by more practice.

7. CONCLUSION

It can be concluded from this research that prediction of the winning team and prediction of the occurrence of the goal can be easily identified by an analysis on the previous data. By using C5.0 algorithm for classification of data set a good accuracy can be achieved for prediction of the occurrence of goal and using Naive Bayes algorithm we can predict the winning team. These analysis can be used for selection of desired player, for making strategy and to give more concentration on week areas of the team to win in the upcoming matches. It also helps sponsors' to make their team more successful and to bid a goog player for their team.

REFERENCES

[1]. https://en.wikipedia.org/wiki/Kaggle
[2]. https://www.kaggle.com/secareanualin/football-events
[3]. https://www.rstudio.com/products/rstudio/
[4]. http://rstudio.github.io/shiny/tutorial/
[5]. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[6]. https://en.wikipedia.org/wiki/C4.5_algorithm
[7]. https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient
[8]. http://www.statisticshowto.com/kendalls-tau/
[9]. https://www.youtube.com/watch?v=V4MgE43SrgM
[10]. https://www.rstudio.com/products/rstudio/features/
[11]. https://www.r-project.org/about.html