Sports Analysis and Outcome Prediction System

Darshan Salian
B.E. Computer Engineering
St. Francis Institute of Technology
Mumbai, Maharashtra, India

Saheel Sawant
B.E. Computer Engineering
St. Francis Institute of Technology
Mumbai, Maharashtra, India

Santosh Maurya
B.E. Computer Engineering
St. Francis Institute of Technology
Mumbai, Maharashtra, India

Abstract

Many efforts has been made in order to predict sports matches result and selecting significant variables. Sports prediction has become increasingly popular in the last few years and many different prediction models have been proposed for evaluating the attributes that lead to lose, draw or win the match. Prediction is very useful in helping team managers, teams and clubs make the right decision to win leagues and tournaments.In this paper, we predict the match outcomes of the matches, by performing a detailed study of past matches and observing the most important attributes that are likely to decide the conclusion.We will be using Multinomial Logistic Regression and Random Forest algorithm to predict the outcome of the matches. Also a comparative analysis is done between Multinomial Logistic regression and Random Forest to find which one of it gives the best accuracy.

Index Terms - Sports, Prediction, Algorithm, Analysis, Machine Learning.

I. Introduction

In today's digitized world, where various sports are being played around the globe, these sporting events are also enthusiastically followed by their supporters. Each of these sports have their specific online platforms, quorums, android applications, etc. wherein the sport enthusiasts update themselves with the specifics of the previous or ongoing games. However, these single sport applications do not offer the feature of outcome prediction or data analysis or basic updates regarding other sports even if a certain user might be interested. Thus, such systems lack a generalized structure. Being restricted to a single sport restricts the scope of the system and makes it inconvenient for the users, where they have switch through multiple platforms to obtain the recent updates of several sports.

Sport prediction is usually treated as a classification problem, with one class (win, lose, or draw) to be predicted. Some researchers, have also looked at the numeric prediction problem, where they predict the winning margin – a numeric value. In sport prediction, large numbers of features can be collected including the historical performance of the teams, results of matches, and data on players, to help different stakeholders understand the odds of winning or losing forthcoming matches. The decision of which team is likely to win is important because of the financial assets involved in the betting process; thus the increasing amount of data related to sports that is now electronically (and often publicly) available, has meant that there has been an increasing interest in developing models and prediction systems to forecast the results of matches.

II. Literature Survey

In paper [1], authors have proposed a logistic regression model to estimate 2015/2016 Barclays' Premier League match results with an accuracy of around 69.5%. They develop this model with the help of data from Barclays Premier League and sofifa.com using four significant variables: Home Offense, Home Defense, Away Offense, and Away Defense. They implement this method in software called Football Predictor. Their work predicts who is going to win a match (home/away), and list out details regarding the odds and probability, International Journal of Pure and Applied Mathematics and the coefficients of regression. This model of just four variables but gives strong prediction accuracy.

In this paper[2], they have proposed method that have been tested on the case of FIFA world cup 2014 tournament. Dataset used was obtained from World cup tournament 'http://www.fifa.com/worldcup/'. Matrix Factorization Model were used to predict the number of goals scored in a match.The prediction algorithm presented in this paper is based only on the base of previous results and does not require any form of match statistics or expert knowledge..The major issue with our this paper lies in the size of the dataset. They used a short-termed dataset, where 64 matches were played in a period of two months, therefore drastic changes in teams' performance (e.g. due to player injuries) were not expected.

In this paper[7], the authors proposed a model to predict the outcomes of football matches in the English Premier League. They trained the final data-set on various machine learning classifiers. The final output or target label is the Full time Match Result (FTR). This label indicates a Home team win (H), an Away team win (A), or a Draw (D).Out of the total of 65 attributes the authors finalized the most relevant 15 attributes.To pickup the most influential parameters the authors used Scatter Matrix.The data is split into training and testing.The authors used Logistic Regression, SVM XG Boost Machine Learning Algorithms to predict the outcome of the matches.

In paper[4], they have then developed a tool called CricAI that can be used to determine the probability of victory in an ODI cricket match given the concerned factors as inputs. The factors being considered for analysis include Home Game Advantage, Day / Night Effect, Winning the Toss. The data set was trained using Naïve Bayes,Decision Trees, Bagging and Boosting. Comparative analysis were done between different machine learning algorithm.

III. Challenges Identified

The certain challenges which the system might have to face are identified while citing the various published papers and other related literature, and are mentioned as follows:

A. Changing Statistics:

Each sport have multiple games across the globe taking place at the same time on a daily basis.. The results of these games will also have an impact on the ranking and statistics of the teams and its players. Therefore, it is vital to take into consideration the updated and correct dataset which will provide proper odds and won't affect the overall accuracy of the system.

B. Internet Connection:

The users must have a strong internet connection in order to access the various features of the system. Absence of which might create a hindrance to smoothly use the system and might lead to technical glitches or improper updations.

C. Relevant Parameters:

To determine the certain parameters among the many taken in to consideration, which has most significant impact on the obtained odds. Thus, it makes it difficult to find better algorithm even after comparative analysis and to predict margin of victory.

IV. Problem Definition

An outcome prediction and sports analysis system for multiple sports rather than focusing on a single sport wherein odds will be presented to users considering various parameters. Users can benefit themselves with the points earned from a correct prediction and other supporting features of the system like recent updates, rankings and analysis and sports quiz.

V. Proposed System Methodology

The prediction and analysis system is explained in two phases as mentioned below:

1. General overview and Block diagram

In this paper, we propose a model to predict the outcomes of various sports. We train the final dataset on various machine classifiers. We compare the performances of each classifier and choose the one that returns the best result. Then, we optimize the classifier that yields the best result to further enhance the model accuracy in making predictions.

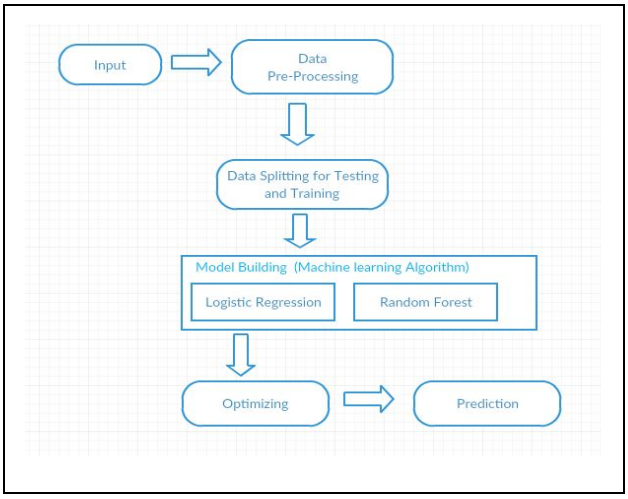

Fig. 01: Basic Flow for Outcome Prediction

A. Dataset Description

We are going to be predicting match outcomes using data from past games for a few seasons. The data-set will be taken from different online sports platforms. Most of the data in the dataset are insignificant and are not needed in our system. Thus we filtered these attributes into a final list through pre-processing which proved to be the most influential for predicting the outcome.

B. Pre-Processing

The data-set we obtained consist of several attributes from each season. A lot of these features are pretty much

unnecessary for making outcome predictions. Hence, our primary task is to clean the data to only retain the features or attributes most. This will help us pick the most influential features that we want to use to build our new data-set. Once we finish building our new set of crucial attributes, we split the data into training and testing data.

C. Model Building

In this module, we finally apply machine learning classifier in order to predict the outcome. We use Logistic Regression: and Random Forest. We compare the result obtained from the above machine learning algorithm and pick the best performing one to give our result.

D. Optimizing :

Finally, we optimize the result by selecting the best performing classifier. We optimize the parameters, to further enhance the performance and accuracy of the model in making our prediction.

E. Prediction :

Thus, in this final step we obtain our target of predicting the outcome of match by selecting the optimized result of the previous step.

2. Proposed Algorithm or Pseudo Code

The two algorithms which will form the basis of the prediction in our system are explained in detail as follows:

Multinomial Logistic Regression:

Multinomial logistic regression is also a classification algorithm same like the logistic regression for binary classification. Whereas in logistic regression for binary classification the classification task is to predict the target class which is of binary type. When it comes to multinomial logistic regression. The idea is to use the logistic regression techniques to predict more than 2 target classes. Since in our output we will be having 3 possibilities i.e Home Win, Away Win and Draw we will be using Multinomial Logistic Regression.

The multinomial logistic regression estimates a separate binary logistic regression model for each dummy variables. The result is M-1 binary logistic regression models. Each model conveys the effect of predictors on the probability of success in that category, in comparison to the reference category. Each model has its own intercept and regression coefficients-the predictors can affect each category differently.

Multinomial Logistic Regression Workflow/ Stages:

- Inputs
- Linear model
- Logits
- Softmax Function
- Cross Entropy
- One-Hot-Encoding

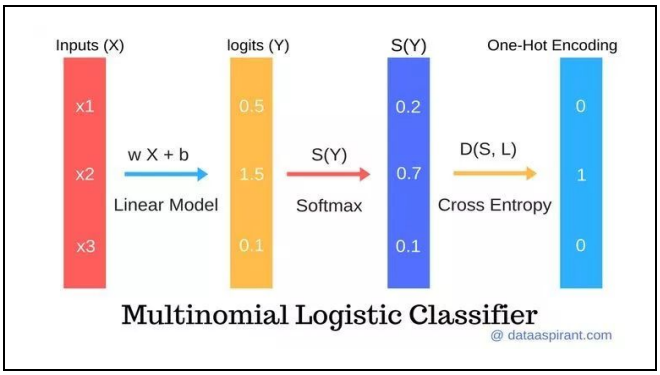

Fig. 02: Execution of Multinomial Logistic Regression

Inputs

The inputs to the multinomial logistic regression are the parameters that we have in the dataset. Since we are going to predict the outcome of matches the parameters will be Home or Away game, Current Team Ranking, Previous Record, Star Player Goals, and other parameters.These parameters will be treated as the inputs for the multinomial logistic regression.

Linear Model

The linear model equation is the same as the linear equation in the linear regression model. Let us consider,

X=[x1,x2,x3], where X is the set of numerical inputs

W=[w1,w2,w3], where W contains input number of weights

The linear model output will be the w1*x1, w2*x2, w3*x3. The weights w1, w2, w3, w4 will update in the training phase.

Logits

The Logits also called as scores. These are just the outputs of the linear model. The Logits will change with the changes in the calculated weights.

Softmax Function

The Softmax function is a probabilistic function which calculates the probabilities for the given score. Using the softmax function return the high probability value for the high scores and fewer probabilities for the remaining scores.

Cross Entropy

The cross entropy is the last stage of multinomial logistic regression. Here it is used to find the similarity distance between the probabilities calculated from the softmax function and the target one-hot-encoding matrix.

One-Hot-Encoding

One-Hot Encoding is a method to represent the target values or categorical attributes into a binary representation.

Random Forest:

Random forest machine learning algorithm were originally proposed by Breiman (2001) and are now a days seen as a mixture between statistical modeling and machine learning. They are an aggregation of a number of classification or regression trees (CARTs). The goal of the partitioning process is to find partitions such that the respective response values are very homogeneous within a partition but very heterogeneous between partitions. CARTs can be used both for metric response (regression trees) and for nominal/ordinal responses (classification trees). For prediction, all response values within a partition are aggregated either by averaging (in regression trees) or simply by counting and using majority vote (in classification trees).

In random forest algorithm, instead of using information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly. Advantages of Random Forest algorithm are there is no need for feature normalization, individual decision trees can work parallely and it is widely used.

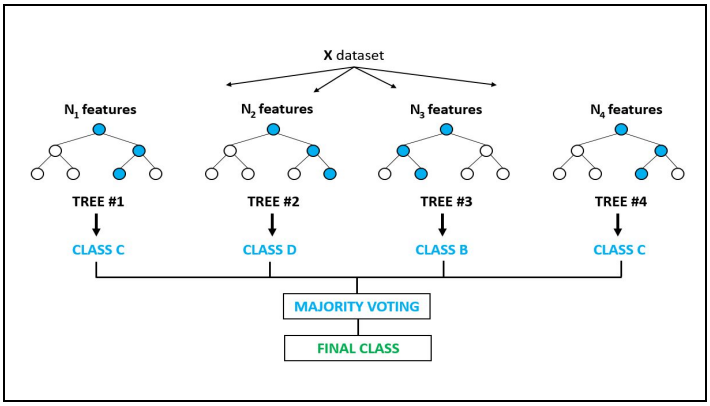

Fig 03: Flowchart for execution in Random Forest

There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage. The whole process is given below.

The Random Forest creation pseudocode is as follows:

1. Randomly select "K" features from total "m" features where k << m
2. Among the "K" features, calculate the node "d" using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat the steps 1 to 3 until "l" number of nodes has been reached
5. Build forest by repeating steps a to d for "n" number times to create "n" number of trees.

In the next stage, with the random forest classifier created, we will make the prediction. The random forest prediction pseudocode is shown below:

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

VI. Performance Evaluation Parameter

Processing Time: To make the system efficient, the processing time should be as less as possible. However, processing time for larger dataset will be more but will give more accurate results.

Accuracy: The accuracy of the odds presented by the system should be as high as possible. This in turn will be helpful in comparative analysis.

Some of the deciding parameters will be alphabetical while some will be numerical. Like for Football, the parameter 'Home Game' will be an alphabetical parameter whereas the 'Number of Goals scored' scored will be a numerical parameter. For the different sports, following parameters will be considered while calculating the optimised prediction odds:

Cricket:

Team Ranking, Record of last five matches, Home or Away game, Result of the Toss , Day or Night game

Football:

Home or Away game, Current Team Ranking, Previous Record, Star Player Goals, and other parameters.

Badminton:

Location (Home Advantage), Court Type, Current Ranking, Game Type (Singles, Doubles, Mixed Doubles), Recent Record.

Kabaddi:

Current Ranking, Recent Record, Star Player Raids, and other parameters.

VII. Experimental setup

Scikit-learn, an open source library which features machine learning models including classification, regression, clustering, etc. will be used. Along with that TensorFlow will be used. It is open source machine learning framework that is used to develop neural networks.

The dataset for other sports can be obtained from sites like ESPN, Cricbuzz, Fotmob, etc.The dataset will be stored in .CSV files.The parameters will be tested using Logistic Regression and Random Forest Algorithm for predicting the outcome.

The data will be divided in training and testing phase.The data for training will be 70 percent and for testing will be 30 percent.

References

[1] Darwin, P & Dra, H. (2016)-'Predicting Football Match Results with Logistic Regression' - International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA).

[2] 'Predicting sports result using latent features'-Stefan Dobravec ,University of Ljubljana, Faculty of electrical engineering, Ljubljana,Slovenia stefan.dobravec@fe.uni-lj.si

[3] Josip Hucaljuk and Alen Rakipović (2011)-'Predicting football scores using machine learning techniques'- Proceedings of the 34th International Convention MIPRO

[4] Alan McCabe and Jarrod Trevathan (2008)-'Artificial Intelligence in Sports Prediction'-Fifth International Conference on Information Technology: New Generations.

[5] Ghada Soliman, Ala'a El-Nabawy, Ahmed Misbah, Seif Eldawlatly (2017)- 'Predicting all star player in the national basketball association using random forest'- Intelligent Systems Conference IntelliSys 2017.

[6] Amal Kaluarachchi and S. Varde Aparna (2010)- 'CricAI:A classification based tool to predict the outcome in ODI cricket'- Fifth International Conference on Information and Automation for Sustainability.

[7] https://acadpubl.eu/hub/2018-118-22/articles/22a/79.pdf

[8] http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

[9] http://dataaspirant.com/2017/03/14/multinomial-logistic-regression-model-works-machine-learning/

FIGURES

Fig. 01: Basic Flow for Outcome Prediction
https://acadpubl.eu/hub/2018-118-22/articles/22a/79.pdf

Fig. 02: Execution of Multinomial Logistic Regression
http://dataaspirant.com/2017/03/14/multinomial-logistic-regression-model-works-machine-learning/

Fig 03: Flowchart for execution in Random Forest
https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57