International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Integrating Feature Selection and Multiclass Classification for Sport Result Prediction

Lydia D Isaac, I. Janani

Abstract: ***Machine learning (ML) has become the most predominant methodology that shows good results in the classification and prediction domains. Predictive systems are being employed to predict events and its results in almost every walk of life. The field of prediction in sports is gaining importance as there is a huge community of betters and sports fans. Moreover team owners and club managers are struggling for Machine learning models that could be used for formulating strategies to win matches. Numerous factors such as results of previous matches, indicators of player performance and opponent information are required to build these models. This paper provides an analysis of such key models focusing on application of machine learning algorithms to sport result prediction. The results obtained helped us to elucidate the best combination of feature selection and classification algorithms that render maximum accuracy in sport result prediction.***

Keywords: Machine Learning, Prediction, Supervised learning, Classification, Feature Selection, Sport Result Prediction

I. INTRODUCTION

Machine learning provides a way for identifying patterns in data and to make use of them to automatically take good decisions or make predictions. One such common task under machine learning, which involves prediction of the target variables in a previously unrevealed data, is termed as classification. The ultimate purpose of classification is to predict something called a class which is nothing but the target variable, by training the classification model with training dataset and then using the same to predict the class with testing data. This type of processing falls under the category called supervised learning as it tries to find out the relationship between input attributes that are otherwise called independent variables and a target attribute that are otherwise called dependent variable. Few applications of classification include medical diagnosis, email filtering, internet traffic interception, click stream analysis and many more.

Prediction of sport result is usually considered as a classification problem, as it would be to predict one from the class: win, lose or draw. It has been always challenging in predicting the outcomes of sport events.

It involves a collection of large number of features like historical data about performance of teams, data about players, results of the matches played and other various data in order to elucidate the odds related to winning or losing the forth coming matches. As the betting process involves huge investments of financial assets, the prediction about the team that is going to win is highly important for the stakeholders involved.

Added to this, sport managers are striving to build appropriate strategies for finding out the appropriate opponent for a team. The tremendous increase in the availability of sports related datasets prove the increase of interest in developing predictive models that could be used for forecasting the results of a match.

In our work, we tried to bring out the best combination of feature selection and classification algorithms that would aptly suit for sport result prediction. This paper would serve best to the research community and the stakeholders related to this application domain. The data sets from 2010 to 2017 are used to construct the model and data set 2018 is used to validate them.

II. RELATED WORK

Multiclass classification is predominantly being used when the classification task involves more than two classes. For example it could be to classify a set of images which may be oranges, apples, or pears. Although there are various classification models available, preprocessing plays a major role in boosting the accuracy of them. The data set may contain irrelevant or unwanted data that do not really contribute to the analysis. Thus dimensionality reduction which is a preprocessing step done before learning is extremely effective in confiscating redundant and unrelated data thereby improving performance of the classifier. However, in terms of efficiency, the increase in dimensionality of data in the recent years pose severe challenges to almost every existing feature selection and extraction methods [10].

Janmenjoy Nayak, Bighnaraj Naik and H. S. Behera, has elucidated the importance of preprocessing. They state that some activities like Data cleaning, Transformation of data and detecting outliers are the important issues to be taken care of for any type of data set since normally, few attribute values cannot be obtained and things to be done on missing values for classification or prediction problems is always been a challenging task [6]. Though there are numerous algorithms in finding out the relationship between the features like Pearson Correlation, Chi-squared, Recursive elimination etc., selecting the features according to the k highest score turned out to be more effective when applied to sport data sets. Many research works has already been done in finding out the results of football matches in prior to the match and also about significant variables to be selected for the same.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Lydia D. Isaac*, Department of Information Technology, Sona College of Technology, Salem, India
Janani. I, Department of Information Technology, Sona College of Technology, Salem, India

One such classification method is Logistic regression. This can be used to predict sports results through regression coefficients. Prediction accuracy (69.5%) of the model built by the authors [12] seemed appreciable. They have used Home Offense, Away Offense, Home Defense and Away Defense as variables.

Support vector machine is another simple algorithm using which a number of research efforts has been done. It is preferred by many data scientists for research as it produces appreciable accuracy with less power consumption. Its objective is to find a hyper plane in an N-dimensional space. Here N denotes the number of data points that distinctly classifies it. Being one among the powerful tools used for supervised learning, SVMs are widely applied to problems involving classification, clustering and regression. Many of its successful applications include plenty of real-world problems like particle identification, text categorization, face recognition, bioinformatics, electrical engineering, civil engineering etc., [6]. It seems that the Learning of SVM depends almost on the training data points count.

Decision tree is a kind of systematic approach for multi class classification. It poses a set of questions to the data set. On the root or parent of each internal node, a question is posted and the data on the node is further split into separate records that have different features. The leaves of the tree refer to the classes in which the dataset is split. The experimental results by the authors [15] have proved Decision Tree to be the most accurate technique in predicting the outcome of football matches with 99.56% accuracy.

KNN (k-nearest neighbor) classifier is the one of the simplest classification algorithm. It does not depend on the data. Whenever a new sample data is found, its k nearest neighbor from the training data is examined. Distance between the two samples may be Euclidian distance or Manhattan distance between their feature vectors [17]. The majority class among the k-nearest neighbors is taken to be the class for the encountered test sample data.

XGBoost is an end to-end tree boosting system that is highly scalable and used widely by data scientists to attain contemporary results on many machine learning challenges [3]. XGBoost system can be applied easily as it is an open source package. Its impact could be found in many machine learning challenges solved world-wide. Enormous solutions to the Kaggle challenges have used this system and have proved to be successful.

Based on all the researches that are being done in this field, we noticed that hybrid models work better in prediction problems. This notion has been applied in the proposed work to find out the correlations between feature selection and classification algorithms.

III. PROPOSED APPROACH

Although there are a number of methods, a structured approach would be useful in obtaining best results in the case of sport result prediction problems. The proposed approach focuses on team sports [1]. The approach is based on the CRISP-DM Framework as shown in Fig.1

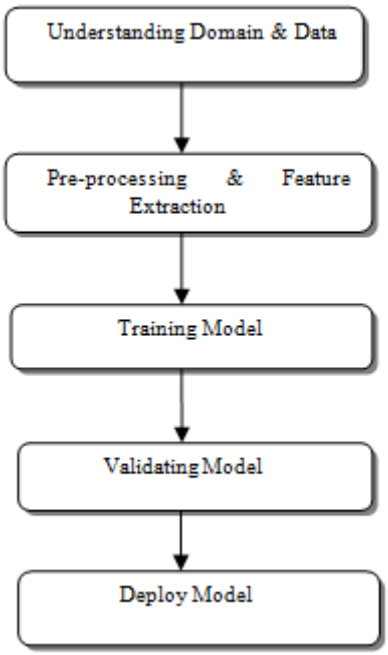

Fig.1 Steps of our proposed Approach

A. Understanding Domain & Data

Domain understanding involves problem analysis identifying the goal of modeling, characteristics of sports and the factors that are involved in determining the outcome of the matches. The objective of the model here is to predict the results of the match just to compete with expert predictions and not for betting. The sports datasets are publically available online. The granularity of the data is to be considered. As this paper is subjected to team sports it may seem unnecessary to see player level data but it could help in understanding whether the presence of a player affects the performance of the team. The proposed work uses 3 class variables (home win, draw, away win).

B. Pre-processing & Feature Selection

The most important step before training a model involves data cleaning and feature extraction as they play vital role in providing accurate results. Dealing with missing data and outliers is necessary to avoid poor predictions. Feature selection is the process by which the features that contribute the most to the prediction variable are selected and thereby remove irrelevant and redundant attributes that deteriorate accuracy of the model.

C. Training Model

There are numerous candidate models that could be used for experimentation. The main three models that we used are Logistics Regression, Support Vector Machine (SVM) and XGBoost. The logistic regression is a way of predictive analysis which is done to obtain descriptions about data and to elucidate the rapport between dependent binary variables and more than one or one ordinal, interval or ratio-level independent or nominal variables. SVMs are models that analyze data used for classification. SVM training algorithm tries building a model that would assign new examples to each and every category. XGBoost is an employment of gradient boosted decision trees which are mainly designed for improved performance and high speed.

Boosting is nothing but an ensemble technique in which new models are being added to overcome the faults made by existing models. This kind of Sequential addition of models will be done until there is no room for further improvements.

D. Validating Model

To evaluate the performance of the model, we classified match results into home win, draw, away win and identified the number of matches the model has correctly predicted comparing it with a standard classification matrix.

Algorithm for Multi Class Classification

Output: Predicted Data for testing Samples
Input: Training Data with Output Labels

1. Begin
2. Import necessary packages and datasets
3. Perform Data preprocessing including data cleaning and feature selection
4. Use Classification Models to predict the results
   4.1 Splitting the data into training and test dataset
   4.2 Using various classification models predict the output label (Home/Away/Draw)

E. Evaluate the accuracy of various classification models.

IV. EXPERIMENTAL SETUP

The aim of the proposed work is to predict the results of a match (win/loose/draw) by coupling the classification models.

1. Import Necessary Packages/Datasets
   The data from 2010 to 2017 data sets were taken for training the model and 2018 data set for testing.

2. Data Cleaning

3. Classification Models to predict match results (Win/Draw/Lose)
Variables used: The Stadium at which is the match is being played (0 –neutral; 1-away team's stadium;2- home team's stadium)
   - Is the match an important match or a friendly match (1- Important, 0 - Friendly)
   - How much the Home team's rank changes compared to the past period
   - Changes in the Away team's rank compared to the past history
   - Ranking Differences between the 2 team's
   - Difference in the 2 team's mean weighted ratings over the past 3 years

4. Classification Models to predict exact goals scored by Home and Away Sides Variables used same as in (3)
5. Visualizing ability/potential of players of the 32 countries
6. Adding variables to build a Poisson model
   Variables used:
   - Average Age
   - Soccer Power Index
   - Average Height
   - Average goals scored per game
   - Total World Cup Appearances
   - Average goals conceded per game
   - Player Potential

7. Predicting World Cup 2018

The results obtained after prediction has been shown in the Table.1.

Classifier	*Training Accuracy*	*Testing Accuracy*
LogisticRegression(Lasso)	***0.58***	***0.48***
LogisticRegression(Ridge)	***0.59***	***0.46***
Support Vector Machine(RBF)	***0.55***	***0.47***
SVM(Linear)	***0.56***	***0.5***
Decision Tree	***0.43***	***0.46***
k-Nearest Neighbours	***0.54***	***0.46***
Random Forest	***0.55***	***0.5***
XGBoost	***0.49***	***0.52***

Table.1. Accuracy percentage obtained

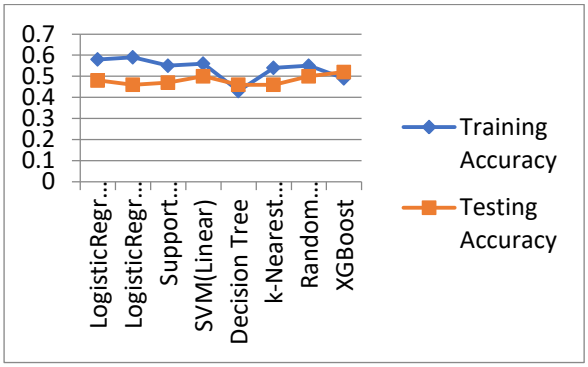

Fig.2. Comparison between various classification models used for sports prediction

V. CONCLUSION

Among the vital applications in sport that necessitates high predictive accuracy is the result prediction of matches. Conventionally, such predictions are done using mathematical and usually statistical models that are normally verified by domain experts. Although there is exponential increase in the use of predictive ML models in the field of sports, still there is a demand for more accurate models. The reason for this is the enormous investments by the betters, and the sport managers longing for some useful knowledge for modelling future match strategies. Most importantly, this paper analyses some of the recent researches rooted on sport prediction that have used multi-lable classification models and feature selection algorithms. Applying the 'SRP-CRISP-DM' framework suggested by P.Bunker and Thabtah[1] the prediction process was carried out for the complex problem of sport result prediction. From the analysis we found that hybrid models could be built for still improving the accuracy of the predictions made.

We found that selection of mRmR(minimum redundancy maximum relevance) features will give high classification accuracy along with possibly reduced feature vector size, where selction of features can be done using genetic algorithms.

REFERENCES

1. Rory P. Bunker, Fadi Thabtah, "A machine learning framework for sport result prediction", Applied Computing and Informatics,Vol.15,Issue 1,January 2019.
2. A. McCabe, J. Trevathan, "Artificial intelligence in sports prediction", Information Technology: New Generations 2008. ITNG 2008. Fifth International Conference on, pp. 1194-1197, 2008.
3. Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", KDD '16, August 13-17, 2016, San Francisco, CA, USA c 2016 ACM.
4. Ramraj S, Nishant Uzir, Sunil R, Shatadeep Banerjee, "Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets", International Journal of Control Theory and Applications, Vol.9,2016.
5. Nicholas I. Sapankevych, Ravi Sankar, "Time Series Prediction Using Support Vector Machines: A Survey IEEE Computational Intelligence Magazine ,Volume: 4 , Issue: 2 , May 2009.
6. Janmenjoy Nayak, Bighnaraj Naik, H. S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges", International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.169-186.
7. Yashima Ahuja & Sumit Kumar Yadav, "Multiclass Classification and Support Vector Machine", Global Journal of Computer Science and Technology Interdisciplinary Volume 12 Issue 11 Version 1.0 Year 2012.
8. Neha Mehra, Surendra Gupta, "Survey on Multiclass Classification Methods", International Journal of Computer Science and Information Technologies, Vol. 4 (4) , 2013, 572 – 576.
9. G. Malik , M. Tarique, "On Machine Learning Techniques For Multi-class Classification", International Journal of Advancements in Research & Technology, Volume 3, Issue 2, February-2014
10. S. Khalid, T. Khalil and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," 2014 Science and Information Conference, London, 2014, pp. 372-378.
11. Cai, Jie & Luo, Jiawei & Wang, Shulin & Yang, Sheng. " Feature selection in machine learning: A new perspective."(2018) Neurocomputing. 300. 10.1016/j.neucom.2017.11.077.
12. D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), George Town, 2016, pp. 1-5.
13. Schumaker R.P., Solieman O.K., Chen H. (2010) Predictive Modeling for Sports and Gaming. In: Sports Data Mining. Integrated Series in Information Systems, vol 26. Springer, Boston, MA.
14. https://hub.packtpub.com/predicting-sports-winners-decision-trees-and-pandas/
15. Che Mohamad Firdaus Che Mohd Rosli et al 2018 J. Phys.: Conf. Ser. 1020012003.
16. Maral Haghighat, Hamid Rastegari and Nasim Nourafza , "A Review of Data Mining Techniques for Result Prediction in Sports", ACSIJ Advances in Computer Science: an International Journal, Vol. 2, Issue 5, No.6 , November 2013.
17. Min-Ling Zhang and Zhi-Hua Zhou, "A k-nearest neighbor based algorithm for multi-label classification", IEEE International Conference on Granular Computing, Beijing, 2005, pp. 718-721 Vol. 2.

AUTHORS PROFILE

Ms. Lydia D. Isaac is currently employed as Assistant Professor in the department of Information Technology at Sona College of Technology, Salem. She has about 11 years of teaching experience and has been a guide for more than 10 UG projects. She owns her Masters' degree in Software Engineering. She has authored and co-authored around 4 publications in International journals and Conferences. Her current area of research involves the domains Machine Learning and Deep Learning in the field of sports. She is a life member of ACM and ISTE.

Janani I is an assistant professor, working in the department of Information Technology at Sona College of technology with 5 years of teaching experience. She had completed M.Tech / Information Technology with the GATE score of 89.5 percentile. Her area of interest is Data Base Management system and Machine Learning in the field of Sports.