ANALYZING IPL MATCH RESULTS USING DATA MINING ALGORITHMS

Shimona.S
Student of Information Technology
MCET,Pollachi
Coimbatore, India
shimonasingam1@gmail.com

Nivetha.S
Student of Information Technology
MCET,Pollachi
Coimbatore, India
nivethashanmugam8@gmail.com

Yuvarani.P
Student of Information Technology
MCET,Pollachi
Coimbatore, India
pyuvarani73@gmail.com

***Abstract*— Cricket is one of the famous outdoor sports that contain a large set of statistical data in real world. As IPL games rise in popularity, it is necessary to examine the possible predictors that affect the outcome of the matches. The article aims at analyzing the IPL cricket match results from the dataset collected (2008-2016) .It focuses on measuring the outcome of Indian Premier League (IPL) matches by applying the existing data mining algorithms to the balanced as well as imbalanced dataset. Oversampling technique is used for imbalanced dataset and then the algorithm is applied. Accuracy is used as the performance metric and calculated by using data mining algorithms. It is also considered as evaluation criteria and percentage will vary according to the different algorithms.**

***Keywords*— CART, accuracy, balanced and imbalanced dataset, oversampling, analyzing, data mining algorithms.**

I. INTRODUCTION

Data mining tools predict the future trends and behaviours, which gives an opportunity to predict the outcome of an IPL (Indian Premier League) match using data mining algorithms. Data mining algorithms have been applied to the IPL dataset and the knowledge from each algorithm has been obtained and analyzed thoroughly as the results are obtained with good accuracy performance. Cricket is one of the most popular sports. The International Cricket Council (ICC) [7] out listed 106 cricket playing nations representing 10 belongs to the full members, 37 of them are associates, and the remaining 59 are considered to be affiliate members.

The Indian Premier League (IPL) [4], sports league was contested during the month of April and May on every year by the teams [5] representing the Indian cities. The result has been predicted using the CART (Classification and Regression Approach) and logistic regression approaches and have analyzed the results of the IPL match using the above approaches. In this paper, data mining algorithms are used to identify the outcome of IPL match for both balanced and imbalanced dataset.

The main motto of the paper is to analyze the outcome of the Indian Premier League (IPL) [4] match. The outcome is analyzed by applying the data mining algorithms to the IPL dataset (2008-2015). Some of the popular variables considered in cricket literature are home-field advantage, coin-toss result, bat-first or second. Thus we measure the outcome of an Indian Premier League (IPL) matches using the data mining algorithms. In this work Classification and Regression (CART) algorithms are implemented to compare the accuracy of the results.

II. ABBREVIATIONS AND ACRONYM

IPL – Indian Premier League; CART – Classification and regression approach; LR- Linear regression; NB- Naive bayes; CSV-Comma separated value; ICC – International Cricket Council; SVM – Support vector machine; SMOTE Synthetic minority oversampling technique.

III. LITERATURE REVIEW

Parag shah [1] in this describes about significant challenges that we face for accurate prediction including the various parameters which affect the outcome of the match. The ball movement gets changed from every over, so it is considered being important to predicting the outcome of each match on every ball. Here they had developed a model that predicts the match result of every ball played. Using Duckworth-Lewis formula the outcome of the match will be predicted for live match. Probability is calculated and figure is plotted for each ball bowled .This model and the probability figure will be very useful for betting industry to decide which team will won the match. Using Par score

concept given by Duckworth &Lewis, probability has been calculated by considering the balls faced, balls left, runs scored, runs left, wicket, wickets left.

H.Ahmad [2] in this, paper explains about the concept of identifying rising stars in cricket domain by using machine learning techniques. Rising stars can be predicted by both bats as well as bowling teams. Distinct features like concept of co-players, team and opposite teams are presented with their mathematical formulation. High accuracy is demonstrated for both robust and statically significant cases. At last the top ranking list of ten rising crickets is compared with International cricket council ranking based on weighted average, performance, evolution and the rising stars scores. Measures are explicitly adopted for rising star prediction in bat and bowling domains. Finally, ranking lists of rising stars based on weighted average, performance evolution and rising star score are presented both domain.

Mehvish Khan, Riddhi Shah [3] in this, the outcome of ODI match depends on various factors. Statistical significance of various variables is explored for this analysis which could explain the outcome of ODI cricket match. The list of key features is home-field advantages, winning the toss, game plan, match type, competing team venue familiarity and season. Logistic regression, SVM, Naïve bayes are there different types of algorithm used for model building. Logistic regression is applied for data that had been already obtained from previous matches SVM and Naïve bayes classifier are used for predictive analysis and model training. It concludes that from 2007, importance of home field advantage on One Day International cricket was statistically studied. It was found that SVM was proved to be a better model based on both the parameters used to predict accuracy and model outcome.

IV. EXISTING SYSTEM

IPL match results [6] had been examined using the existing data mining algorithms for the dataset collected during the year of 2008 – 2016. In this dataset, there are lot of missing information and lacking attributes. Hence they do not bother about the imbalanced dataset and had applied the algorithms for analyzing the results. The drawback of this system leads to the misleading accuracy performance for the imbalanced data when compared to balanced data.

V. PROPOSED SYSTEM

The IPL [4] is one of the most-attended cricket league in the world and it ranks sixth position on all sports leagues. Therefore the proposed system focuses on analyzing the IPL matches results [6] by applying classification algorithm in data mining. Results must be greater than the existing system and accuracy will be compared to the above algorithms. The output will be of calculating the accuracy for the IPL match, after oversampling the imbalanced dataset.

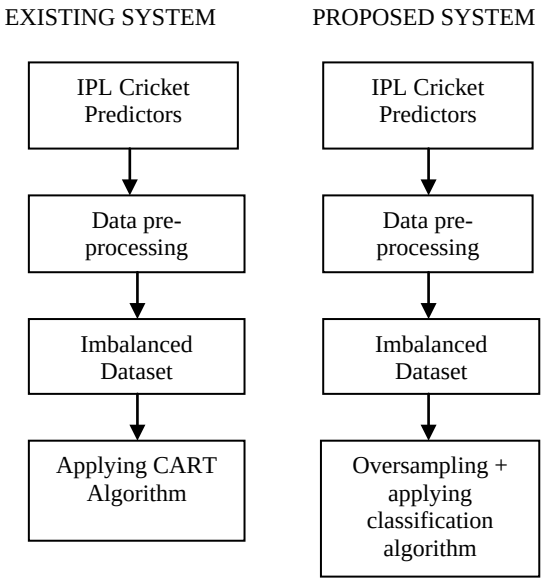

Fig. 1. Architecture Comparison

Data set [8] has been collected for the IPL Match (season from 2008-2016) in Kaggle website. It comprises of 12 attributes and 578 entities which are in Comma Separated Value (CSV) format. The module contains the following steps,

Data cleaning is process of removing the incomplete data, missing information, and also detecting the inaccurate records and replacing it with correct records.

Data transformation is referred to converting one form of data into another form of data. It is considered to be both simple and complexes based on data between initial data and final data.

Data pre-processing is a technique that involves transforming the inaccurate form of data onto accurate form. Real world data is often represented to be in form of inconsistent and inappropriate records. Attributes which are not necessary for the prediction of the match have been excluded from the dataset. The removed attributes are id, player of the match, result, toss decision, and season. The final attributes are taken for prediction includes city, team1, team2, toss winner, win_by_run, win_by_wicket and winner. The missing data or information on the dataset is neglected and the number of entities has been shortlisted to 524.As we are going to process the dataset in the Matlab, we have to replace the character data to numeric data. The team names are replaced and unique number has been allocated to each team.

As we are going to predict the results of the match, we have to create the test dataset with preprocessing. Training dataset has to be created in this process. It includes attributes of team1, team2, toss winner, win_by_run, win_by_wicket, and winner. The attribute winner is partitioned in training set to foresee the result of the match.

A. APPLYING CART ALGORITHM FOR IMBALANCED DATASET

The existing algorithms that are applied to the IPL dataset are,

i. *Decision tree:*

It is used as supervised tool for field of data mining. It consists of two processes namely training and testing. Interpretation phase of decision tree is composed of deficiencies in neural networks. Therefore tree bagging has been implemented in decision tree process. Bootstraps aggregation is represented as ensemble method of decision trees. Each tree is grown as independent drawn bootstrap of input data. Errors are referred to as “out of bag” which is not included in this replica. Tree bagging takes as leverage of predictions for predicting the unseen data.

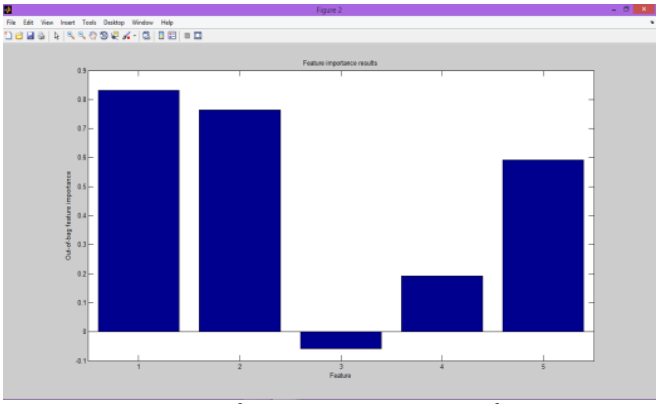

Fig. 2. Bar graph representing increase in prediction error

ii. *Naive Bayes:*

Naive bayes classifier is used where one class is independent of another class. Thus normal and kernel distribution were implemented in our paper.

a) Normal distribution

It is applicable to predictors that have normal distribution of each class. Here the naive bayes classifier estimates a separate distributed over each one by computing mean and standard deviation from training data.

b) Kernel distribution

It is appropriate for the predictors which will have continuous distribution. Here naive bayes compute a separate kernel of each class based on the training data.

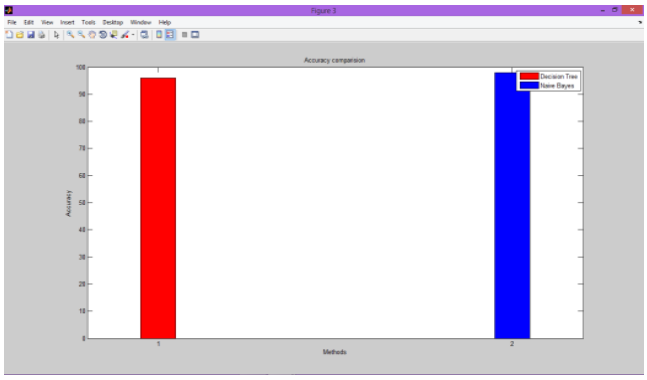

Fig. 3. Accuracy comparison of decision tree and naive bayes

iii. *Regression*

A regression method is used where the target values is known. Regression algorithm is used for estimating the values of target as functions of predictors while building the training process. The relationship between the target and predictors is summarized and can be applied to different data set in which the target values are known.

These models can be tested by the various statics that measure the difference between accepted and predicted values. The procedure gets executed in the respective dataset and the result is divided into two processes, Actual and Predicted items. Actual items refer to the winner of the match in the dataset and predicted item represents the execution of the algorithm and its driven results.

B. OVERSAMPLING APPROACH FOR BALANCING THE DATASET

While performing the analysis in well balanced data set, many of the classifiers seem to be well working in response variable of dataset. Complex situations arise when the dataset is imbalanced. So oversampling technique is used to adjust the class distribution of a dataset. By using oversampling we can change the class distribution of the training data. The reason for altering the class distribution is for learning with highly skewed datasets to impose the non uniform misclassification costs. Oversampling is considered to work well in improving the classifications for imbalanced dataset using the decision tree or other classifiers. When dealing with the imbalanced data sets, data mining algorithms for difficulties such as the predictions estimated are biased and of misleading accuracy. This mainly occurs due to the lack of information about the minority class. Data mining algorithms assume that the dataset is balanced and therefore classify every test case sample of minority class to improve the accuracy metric. To overcome this issue, sampling techniques has been considered as a solution.

To balance the dataset we have to mainly follow three procedures,

a) Undersampling the majority class
b) Oversampling the minority class

a). UNDERSAMPLING THE MAJORITY CLASS

The most common and simplest solution to handle imbalanced data is to under sample the majority class. Undersampling is used to increase the sensitivity and to solve the imbalance issue. A reason for this could be of training the classifiers using few samples.

b). UNDERSAMPLING AND SMOTE COMBINATION

Majority class has been under sampled by randomly removing samples of majority class until minority class becomes the specified percentage of majority class. The terminology explains if we under sampling the majority class of 200% means then modified dataset will have twice as many elements from the minority as from the majority class. As the minority class has 50 samples and majority has 200 samples, when we perform under sampling majority of 200% means it will end up having 25 samples. By applying the combination of both under sampling and oversampling, the initial bias towards the majority class is reversed in favour of minority class.

ADVANTAGES UNDERSAMPLING USING SMOTE

Smote is a method to generate synthetic samples that can potentially eliminate the class imbalanced problem. Here we have applied to smote the high dimensional class imbalanced data and also used some results to explain the behavior of smoting. The main findings of our analyses are,

1. In the low-dimensional set SMOTE is efficient in reducing the class-imbalance problem with most classifiers.
2. When data are high-dimensional SMOTE is beneficial of classifiers if variable selection is performed before SMOTE.
3. Undersampling or, for some classifiers, cut-off adjustment are preferable to SMOTE for high-dimensional class-prediction tasks.

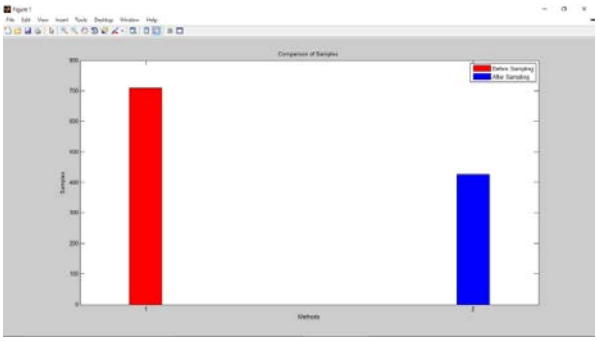

Fig. 4. Comparison of samples

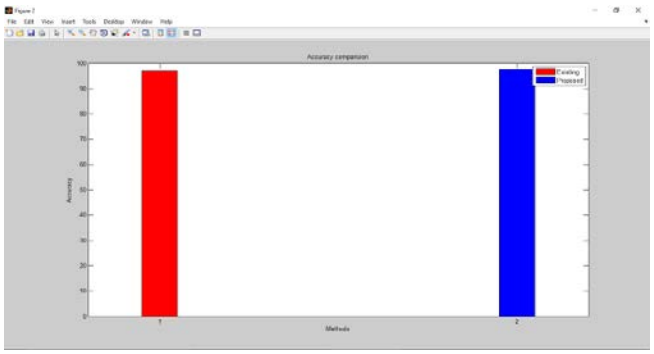

Fig. 5. Accuracy comparison of balanced and imbalanced dataset

C. OUTCOME AND EVALUATION

The existing algorithms have been applied to the imbalanced dataset and results are estimated. From the results obtained, comparison of each algorithm is based on the accuracy calculated. When applying these algorithms to the IPL dataset, accuracy has been calculated and best algorithm is chosen based on it. Evaluation criteria are represented as the accuracy rate of the model on the test dataset. The table below gives the accuracy rate of each algorithm along with how it performed using percentage split method for the imbalanced data. In comparison to three different classification and regression algorithms, we finalize that the naive bayes has the best accuracy and it depict the outcome of the match accurately.
The accuracy results for the three algorithms are,

S.NO	Accuracy rate of the Algorithms	
	Name of the Algorithms	***Accuracy Percentage***
1	Decision Tree	96.01%
2	Naive Bayes	96.92%
3	Linear Regression	93.30%

TABLE I. ACCURACY RATE OF THE ALGORITHMS

Balanced dataset provides a better classification when compared to imbalanced dataset. Therefore, the classification algorithm (Naive bayes) is applied to the balanced dataset and the results are obtained using oversampling.

S.NO	Accuracy rate of the both balanced and imbalanced dataset	
	Algorithm(Naive Bayes)	***Accuracy Percentage***
1	Before Sampling	96.98%
2	After Sampling	97.56%

TABLE II. ACCURACY RATE OF THE BOTH BALANCED AND IMBALANCED DATASET

VI. RESULTS AND DISCUSSIONS

The main objective behind the need to pre-process the imbalanced data is typically the classifiers are more sensitive to detect the majority class and less sensitive to minority

class. For making the dataset as balanced two approaches are followed, such as oversampling and undersampling. Here we conclude that the balanced dataset will provide the greater accuracy than the imbalanced while analyzing the outcome. Hence, the use of sampling on imbalanced dataset is justified by above acquired results.

VII. CONCLUSION

This paper has intended on analyzing the results of the IPL match during the year 2008-2016 by applying the data mining algorithms on both the balanced as well as imbalanced dataset. The model which is used to analyze the results of matches was built successfully with accuracy rate of 97% for the balanced dataset using the classifiers (i.e) after oversampling the imbalanced IPL dataset [8]. The outcome values are higher than the imbalanced dataset and errors are lesser.

ACKNOWLEDGMENT

The authors would like to thank the guide Ms.T.Sumathi, M.Tech, Assistant Professor, Information Technology for bringing this paper to its logical conclusion. This work was supported and assisted by Dr. S. Ramakrishnan, M.E., Ph.D., Professor and Head of the Department, Information Technology at Dr.Mahalingam college of Engineering and Technology, Pollachi.

REFERENCES

[1] Parag Shah, “Predicting Outcome of Live Cricket Match Using Duckworth-Lewis Par Score”, Publisher: International Journal of Latest Technology in Engineering, Management & Applied Science, Volume VI, Issue VIIS, July 2017.

[2] Haseeb Ahmad, Ali Daud, Licheng Wang, Haibo Hong, Hussain Dawood, and Yixian Yang , “Prediction of Rising Stars in the Game of Cricket”, Publisher: IEEE Access, Issue March 4 2017.

[3] Mehvish Khan, Riddhi Shah, “Role of External Factors on Outcome of One Day International Cricket (ODI) Match and Predictive Analysis”, Publisher: International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2015.

[4] https://en.wikipedia.org/wiki/2016_Indian_Premier_League

[5] https://www.sportskeeda.com/cricket/ipl-2016-teams-full-players-list

[6] http://statisticstimes.com/sports/all-ipl-points-table.php

[7] https://www.icc-cricket.com/rankings/mens/team-rankings/test

[8] https://www.kaggle.com/manasgarg/ipl