IPL DREAM TEAM - A PREDICTION SOFTWARE BASED ON
DATA MINING AND STATISTICAL ANALYSIS

Mrs. Jayshree Hajgude 1, Aishwarya Parameshwaran 2, Krishna Nambi 2, Anupama Sakalkar 2, Darshil Sanghvi 2

1Assistant Professor, 2B.E. Student, Department of Information Technology, V.E.S Institute of Technology, Mumbai, India

ABSTRACT

The Indian Premier League (IPL) is an annual Indian Twenty20 cricket tournament, founded in 2008 by the BCCI. Management of an IPL team requires handling of a large amount of data pertaining to the players and team management and the probabilistic values associated with the practical use of the data. Our software, the IPL Dream Team is developed for use by the owner of an IPL team franchise. The software we propose aims at providing a comprehensive solution to the handling of all such complex data and presenting it to the user in a clear, concise and readily usable form thereby eliminating the expenses involved in manual data management.

Keywords: IPL, Naive Bayes, Sentiment Mining, Data Mining, Statistical Analysis

[1] INTRODUCTION

The Indian Premier League (IPL) is an annual Indian Twenty20 cricket tournament, founded in 2008 by the BCCI. It consists of a number of teams (currently 8) which are named after Indian cities and owned by franchises, and is held in India between April and June. The first part of the league comprises of live auctions which involves the participation of several team owners. Currently there are no popular assistors readily available to aid the owners in making their decisions during the same.
This drove our team towards developing a solution to help the owners by providing them with a tool that makes intelligent suggestions for

1. which players to replace
2. which players to buy

The motive behind developing a viable prediction software is elimination of the efforts and amount of human expertise required to make beneficial decisions. The software takes into consideration the past performance of each player in the league as well as his popularity among the fans for predicting if a player would be acceptable to the owner. This mainly enables the owners to optimize their choices and automate the team prediction and formation process.

[2] FEATURES

The developed project incorporates the following features:

1. The software would be customized accordingly for every team in the premiere league. The home page consists of the owner’s team profile, photo stream and it will also list the current popular players of the owner’s team and live updates about the team from their very own twitter handle.
2. Sentiment mining has been implemented by gauging the popularity of every player to determine which player could be projected as the face of the team in publicity campaigns and to incorporate the results of the findings in calculating the overall rating of every player for the formation of the ultimate dream team before the onset of the auction.
3. Web mining is performed to obtain statistics for every player.
4. Every source site from which data has to be obtained has been checked for its authenticity in order to filter out the unreliable data and only the reliable sources has been used for obtaining statistics.
5. The software consists of a feature called friendly negotiations wherein the owner of the IPL team can negotiate with the owner of other IPL teams for an exchange of player by comparing their statistics and market price value so that the franchise can generate revenue by swapping big players or strengthen the team by swapping weak players in the squad. This feature can be used by the team owners before the onset of the auction session.
6. The software also consists of a feature called portfolio where the performance statistics of every player over the years are displayed in a graphical form as well as in a tabular format which helps the owner to examine his teams’ statistics over the years.
7. After the generation of a list of acceptable players, the software also allows the user to input his experience to add another dimension to the data and performs computations with the added dimension as well. These dimensions are accepted in the form of statistics filters where the owner can set a range of values for statistics such as batting average, bowling economy, international strike rate, etc. which further filters the list generated by the prediction algorithm. On obtaining the statistics, a prediction algorithm would be used to determine which players would be best suited for the team on the basis of some acceptance parameters set for the owner.
8. The list generated after the filtering done by the owner’s parameters will undergo sentiment mining which will rank the players based on their popularity and a final list of top 20 players will be generated as the "Dream Team" for the owner as he visualizes it, before the auction.

9. The Auction tab will display the annual budget of the owner’s franchise for the current season’s auction session as well as the final Dream Team list along with the teams’ current players to be retained for the season. Using the list of players generated by the software to be auctioned this season along with their base price, the user would be able to input his bidding amount after he wins the bid for a particular player during the auction session and the final list of bought players will be displayed. The owner can buy players by referring to his Dream Team list and generate the final Dream Team as he visualizes it.

[3.1] SENTIMENT ANALYSIS BASED ON TWITTER DATA

![Figure 2. The workflow for Adaptive Sentiment Analysis enforced](image)

The above defined basic steps obtained by the study of several papers, have been comprehended and optimized to fit the requirements of our software\(^3\)[4][5][6].
The first step involves extracting the Tweets using the Twitter Search API and identifying the features within it\(^4\). The features mainly comprise of:

- Usernames
- Hyperlinks
- Smileys
- Text (Parts of Speech)

Identifying the redundant features aids us in eliminating them at a preliminary stage, thus improving the performance of the software. Usernames and hyperlinks, in our context, are unnecessary and hence, are eliminated using basic regular expressions. Also, dictionaries for different parts of speech are used in order to eliminate the unnecessary conjunctions, prepositions and articles.

The next step involves comparison of the elements in the tweet list with a unigram lexicon\(^3\). The tweet list has been obtained by splitting the tweet stream word-by-word. The tweet stream is acquired and written using JSON. JSON being smaller is easier to parse and human-readable. The lexicon has a polarity associated with each term in it based on its positivity or negativity in relation with a set number of positive and negative instances indicated in the file as well.

The next step is the most important part of the algorithm since it involves defining the polarity bands. Defining a fixed baseline of 33% of the total possible instances is the generally accepted principle while classifying into three labels- Positive, Neutral and Negative\(^3\). But the possible positivity or negativity of the tweets may differ according to a number of factors:

1. The perception of the individual tweeting
2. The number of relevant elements in the tweet-list
3. The steepness of the Neutral band

We fail to consider the possibility that the positivity or negativity of a term may differ according to how often or in what context a person tends to use it. So the polarity of a term technically differs with the frequency of usage of the term by an individual. We intend to take this factor into consideration in the future. Currently we are focusing on the second and the third factor on the above list.

If after preprocessing there are only 50% relevant elements, having fixed baselines for the bands can prove to be erroneous. So according to our algorithm the baselines alter in coordination with the changing size of the tweet-list. This makes the algorithm more adaptive and efficient.

Linear regression is enforced on the polarity of individual terms to obtain and net polarity. This net polarity is then fit within the band-limits of a label. Given below is a detailed instance of the enforced algorithm. This algorithm was devised by us keeping in mind the fluctuations in the number of relevant popular tweets.

Algorithm:

1. Extract tweets using Twitter Search API and write it to a file > tweets.txt
2. Eliminate the usernames, hyperlinks, prepositions, conjunctions and articles and write individual words into an array.
3. Now compare individual words/terms with the terms in the lexicon and derive their polarity.
4. To obtain the net polarity perform the necessary computation as follows:
   \[
   \text{Net polarity} = 0 \\
   \text{For all terms in tweet-list} \\
   \text{Net polarity} = \text{Net polarity} + \text{Polarity of term in consideration}
   \]
5. Define Positive, Neutral and Negative bands as follows:
   5.1 Set the lower-limit for Negative band as \(\text{tweetlist size} \times \text{(lowest possible polarity)}\)
   5.2 Set the upper-limit for Positive band as \(\text{tweetlist size} \times \text{(highest possible polarity)}\)
   5.3 The negative band extends from lower-limit to \([\text{lower-limit} + (48\% \text{ of total range})]\)
   5.4 The positive band extends from \([\text{upper-limit} – (49.75\% \text{ of the total range})]\) to upper-limit
   5.5 The remaining portion of the range is the neutral band
6. Fit the net polarity into a band according to its boundaries

The limiting percentages 48 and 49.75 have been determined based on repeatedly running the algorithm on pertinent sample text with varying standard deviations starting with a 33\% baseline for the end limits with a deviation of 1\%. The sample data used for the study is acquired from reliable sources\[8\]. Repeatedly running the algorithm and altering the deviation for the bands expanding and narrowed the bands, eventually giving the probabilistically correct band limits with the accuracy rate of 96.74\%.

**[3.2] PREDICTION AND FILTERING**

The software now forms the dream team with the help of Bayesian prediction technique and parameter-based filtering. The main database consists of statistics of all players currently playing in IPL. Two separate data sets are required to train the algorithm for batsmen and bowlers separately, since batsmen and bowlers are compared using different parameters. The training dataset has been built from the original dataset using replacement strategy and has been tested by partitioning the main database into batsmen and bowlers. When the owner asks for a list of Acceptable players to be generated, PHP code is used to access the main database. The code scans through the main database, identifying a player as a batsman or a bowler based on the 'Position' attribute. When a player is identified, the algorithm compares certain fixed statistics of the player with all the rows of relevant training dataset and calculates the probability of whether a player is acceptable or not. The player is classified as Acceptable or Not Acceptable depending on which class's probability is greater.
A list of all players classified as Acceptable is returned. PHP code displays the output using dynamic tables.
IPL DREAM TEAM - A PREDICTION SOFTWARE BASED ON DATA MINING AND STATISTICAL ANALYSIS

Figure 3. The datasets required to train the prediction algorithm

Naive Bayes generates Acceptable players from the entire player pool. Once the owner of the team has the list of Acceptable players, it follows that he would want to select some players for his own team based on his own range of parameters such as Average, Strike rate etc. Filters have been implemented for the owner to input his own parameters in further refining the list of batsmen and bowlers using jQuery and Javascript. When the owner asks for players to be filtered based on his parameters, php code is used to access a MySQL database. Statistics of every player from the list of Acceptable players are extracted from the MySQL database. Players which fit in the given range of parameters are compiled by the php code and displayed to the user using dynamic table. Once the filtered list is generated, the list undergoes another refinement in terms of sentiment mining. Sentiment mining determines the most popular players from the list, thereby adding another parameter for the owner to select the most appropriate players. Finally, after the players have undergone sentiment mining, the dream team is formed.

![Table showing statistics of players](image)

**Figure 4. Filters implemented to enable owner to set his own parameter values**

In order to keep a track of players being bought and sold during the auction, the auction page contains players divided into four categories: Players of the owner's team, Players of the owner's Dream Team, Players to be auctioned and Players bought by the owner.
Players of the owner's team are filled in automatically. Players of the Dream Team are obtained after the owner applies all filters provided. ‘Players To Be Auctioned’ is the list of all Acceptable players obtained after applying Naive Bayes algorithm.

Once a player has been bought the owner can enter the amount for which he was bought, which causes the player to be transferred in the Bought list thereby removing him from the To Be Auctioned list.

[4] CONCLUSION

Thus from the performed survey and research, we refurbished a model to mine relevant data to be used by prediction algorithm in order to obtain the statistical analysis of each player. This will in turn helped in determining the players that could be sold or bought in the auction and also assisted in evaluating the performance of individual players for the further seasons to come. Also we enforced a novel adaptive sentiment analysis methodology based on regression in the light of the research conducted on the existing techniques. This in turn helped us develop a viable and highly efficient solution aiding the IPL owners in their decision-making process.

REFERENCES

[1] Indian Premier League, from Wikipedia, the free encyclopedia
[2] Indian Premier League Official Website
   http://www.iplt20.com
[8] Opinion Mining, Sentiment Analysis, Opinion Extraction
   http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.htm

Mrs. Jayshree Hajgude is an Assistant Professor in the Department of Information Technology at Vivekanand Education Society’s Institute of Technology, Mumbai, India.

Aishwarya Parameshwaran, Krishna Nambi, Anupama Sakhalkar and Darshil Sanghvi are undergraduate students currently pursuing their Bachelor of Engineering degree in the field of Information Technology at Vivekanand Education Society’s Institute of Technology, Mumbai, India. They will graduate in 2015.