ICC T20 Cricket World Cup Prediction Based Data Analytics and Data Mining Technique

Anubha Roy

Sakshi Pandey

Priyanka Bhatia

PROF. M. A Rane

Department of Information Technology Bharati Vidyapeeth's College of Engineering for Women, Katraj, Pune, Maharashtra, India.

Abstract:

With the advent of statistical modeling in sports, predicting the outcome of a game has been established as a fundamental problem. Cricket is one of the most popular team games in the world. We embark on predicting the outcome of a One Day International (ODI) cricket match using a supervised learning approach from a team composition perspective. Our work suggests that the relative team strength between the competing teams forms a distinctive feature for predicting the winner. Modeling the team strength boils down to modeling individual player's batting and bowling performances, forming the basis of our approach. We use career statistics as well as the recent performances of a player to model him. Player independent factors have also been considered to predict the outcome of a match. We will show that the k-Nearest Neighbor (KNN) algorithm yields better results as compared to other classifiers like Naïve Bayes, Support Vector Machine (SVM), etc. The performance is affected by the type, size and quality of the data.

Keyword:

Melanoma; Data Mining, Machine Learning, Kth Nearest neighbor, Naïve Bayes.

Introduction:

Statistical modeling has been used in sports for decades and has contributed significantly to the success of the field. Cricket is one of the most popular sports in the world, second only to soccer. Various natural factors affecting the game, enormous media coverage, and a huge betting market have given strong incentives to model the game from various perspectives. However, the complex rules governing the game, the ability of players and their performances on a given day, and various other natural parameters play an integral role in affecting the outcome of a cricket match. This presents significant challenges in predicting the accurate results of a game.

The game of cricket is played in three formats - Test Matches, ODIs and T20s. We focus our research on ODIs, the most popular format of the game. To predict the outcome of ODI cricket matches, we will propose an approach where we first estimate the batting and bowling potentials of the 22 players playing the match using their career

statistics and active participation in recent games. We will use these player potentials to render the relative dominance one team has over the other. Taking two other base features into account, namely, toss decision and the venue of the match, along with the relative team strength, we adopt supervised learning algorithms to predict the winner of the match. The major algorithms used in the project will be:

Support Vector Machine:

SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

Euclidean Distance(x, xi) = sqrt(sum((xj – xij)^2))

Naive Bayes:

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. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: 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.

Bayesian classifiers are based on Bayes' theorem.

Bayes Theorem: Let X be a data tuple and C be a class label. Let X belongs to class C, then,

P(C|X) = P(X|C)P(C) / P(X) where;

• P(C|X) is the posterior probability of class C given predictor X.

• P(C) is the prior probability of class.

• P(X|C) is the posterior probability of X given the class C.

• P(X) is the prior probability of predictor.

The major contributions of the paper will be:

- We will propose novel methods to model batsmen, bowlers, and teams, using various career statistics and recent performances of the players.
- To predict the winner of ODI cricket matches, we propose a novel dynamic approach to react to the changes in player combinations.

We will calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction. For this we have to convert the data set into a frequency table.

Related work:

Better predictive modeling depends on a better understanding of the data and attributes selection. We have to choose between some data mining algorithm. We have chosen data mining as it is very flexible in predictive modeling. [5] Prediction, when the game is in progress, is a tough task and it needs depending on the best attributes

that influence the match outcome. Such solutions designed for offline usage and no in-game effects were taken care of. There have been some recent works (20) about in-game decision making to find how much time remaining in the game without making any prior prediction model. There were several works done in cricket. Bailey and Clarke and Sankaranarayanan [2] used a machine learning approach to predict the result of a one day match depending on the previous data 14 and in-game data. Akhtar and Scarf used multinomial logistic regression in their work on predicting an outcome of test matches played between two teams. Choudhury [1] used Artificial Neural Network to predict the result of a multi-team one-day cricket tournament depending on the past 10 years data. They used a training set to model the data in a neural network. Again there were no in-play effects that were taken care of. For baseball, Ganeshapillai and Guttag developed a prediction model that decides when to change the starting pitcher as the game progresses. [4] It is very much similar to our workflow, where they used the combination of previous data and in-game data to predict a pitcher's performance. Tulabandhula and Rudin were designed a real-time prediction and decision system for professional car racing. The model decides when are the best time for a tire change and how many of them. These works supplied huge encouragement and informative ideas in our research.

Motivation:

Though Sports Analytics has shown a lot improvement and advancement but still this interesting field has been lagging in terms of applications in real sports. Many times news channels organize debates on predictions of cricket matches. Sponsors and Businessmen invest a lot of money on teams without knowing whether their team will win or not. Using Machine Learning if we are able to predict the winning team, then it will be easy for the sponsors and other investors to think whether they should sponsor the team or not. It will also become easy to find the places where the team is lagging and this will help the team to work on the particular areas.

System Architecture:

The project is the developed module that has a User login registration and Admin login Registration. The system finds the Generated Winning team by using classification when a user-provided dataset. Admin adds all the information related team player as well as a Team also. The Proposed System is Find out the Predicated Winning Team and generate Result. This system work on two-way working client-server

Generation.

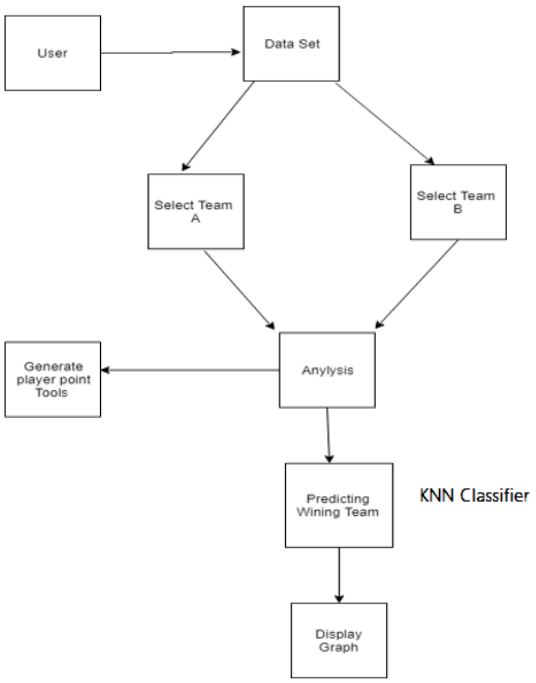

Fig. The Proposed System

Fig. The Proposed System

1. Registration: The user can register by using basic information (First Name, Last Name, Email ID, Password Phone Number, etc.)

2. Login: After Registration, the user can log in.

3. Classification: Select the team and predict the winning team by player runs.

4. Solution: After classification predicts the winning team.

Conclusion:

The paper will conclude the problem of predicting the outcome of an ODI cricket match using the statistics of 366 matches with the help of KNN classifier and Naive Bayes algorithm. It will devise method to find the cricket match outcome prediction, team structure analysis and player recommendation system using the statistics of the players extracted from a particular tournament. We will observe that simple features can yield very promising results.

Reference:

[1] A. Aburas, Machine Learning Algorithms for Big Data Project, Durban: University of Kwa-Zulu Natal, 2018.

[2] Jesus Maillo, Sergio Ramírez, Isaac Triguero & Francisco Herrera, kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data, Knowledge-Based Systems 117 (2017) 3–15.

[3] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, Edin Muharemagic, Deep learning applications and challenges in big data analytics, Journal of Big Data volume 2, Article number: 1 (2015)

[4] Weiping Cui, Lei Huang, A map reduce solution for knowledge reduction in big data, International Journal of Computer Science and Applications, Vol. 13, No. 1, pp. 17 – 30, 2016

[5] Karl Weiss, Taghi M. Khoshgoftaar, DingDing Wang, "A survey of transfer learning",Weiss et al. J Big Data (2016) 3:9.