Winning Prediction in One Day International (ODI) Cricket using Machine Learning Techniques

[1]Abhishek Naik, [2]Sahil Mulani, [3]Minakshee Naik, [4] Shivanee Pawar,
Computer Department
AISSMS IOIT, Pune, India

Abstract: Cricket prediction can be viewed as one of the objectives of sports analytics, which aims at helping decision makers to gain competitive advantage. The difficulty of this task depends on many factors, like the availability of data for the past events, the ability to gather data for future events, the knowledge needed to interpret gathered data, and others. Various techniques for modeling a cricket match exist that yield different result prediction algorithms. The modeling can be put under the four generic categories: empirical models, dynamic systems, statistical techniques and artificial intelligence (including expert systems). In the artificial intelligence category, there are several approaches that focus on Bayesian network modeling. The Matrix factorization technique became very popular in the field of multimedia content recommender systems where it showed good scalability and predictive accuracy. The idea behind using the latent features in our case is to be able to build a successful model over existing matrix factorization.

Keywords*-Neural Network, Linear Regression, Clustering, Decision Tree.*

I. INTRODUCTION

Cricket prediction can be viewed as one of the objectives of sports analytics, which aims at helping decision makers to gain competitive advantage. The difficulty of this task depends on many factors, like the availability of data for the past events, the ability to gather data for future events, the knowledge needed to interpret gathered data, and others. Various techniques for modeling a cricket match exist that yield different result prediction algorithms. The modeling can be put under the four generic categories: empirical models, dynamic systems, statistical techniques and artificial intelligence (including expert systems). In the artificial intelligence category, there are several approaches that focus on Bayesian network modeling. The Matrix factorization technique became very popular in the field of multimedia content recommender systems where it showed good scalability and predictive accuracy. The idea behind using the latent features in our case is to be able to build a successful model over existing matrix factorization.

Objective of this report is to prepare a model for predicting outcome of One Day International Cricket match and to measure accuracy of model on the basis of historical data. We all know that ODI Cricket is very popular. What makes ODI so popular is not only its shorter duration but also unpredictability. Only a single player with the good hitting of 10 overs can change the mood of a game unlike test cricket, where to win game team has to dominate days not some
overs. Not only temperament, calmness, experiences but aggression and hitting matters a lot and this all make ODI most unpredictable format of the game. Just to support my point further, Bangladesh has won 101 ODI yet but could only manage to win 8 tests so far and all these 8 win comes in recent time. You can expect teams like Ireland, Afghanistan or Canada to win an ODI game but never in a test match. Unpredictability makes ODI most popular format of cricket. Although unpredictable but the outcome of a game depends on a lot of factors like toss, venue, weather conditions, ground previous results statistics, players experience and current form and team's previous results against the opponent. So it would be nice to use these parameters and to see up to what extent each particular parameter can affect the outcome and to prepare a model, which can predict the outcome of a game.

II. ALGORITHM

Decision Tree Algorithm:
In decision tree, a tree like model containing decision and answer node is built. Entropy is calculated to partition data into subsets containing instances with similar values and then the root node is decided based on the attribute that returns the highest information gain. Sets of rules are mapped from root node to leaf node.

Information Gain: After splitting the datasets, info gain is calculated based on the change in entropy. Information gain is calculated as follows:

Gain [T, X]. Entropy [T]-Entropy [T, X]

Initially the entropy of the parameters is calculated. Then, information gain is calculated based on above given formula. The attribute having the maximum information given is chosen as a root. Entropy 0 indicates a leaf node where as entropy greater than 0 requires splitting. Predicted class is given by the leaf node.

1. Select the best attribute from the dataset and place it at the root of the tree.
2. Break down or split the training datasets into multiple subsets.
3. Repeat step one and two on the respective subsets unless and until you find the leaf nodes of the tree.

Attributes Selection

Suppose the datasets include 'n' number of attributes then it becomes complicated to decide which attribute should be placed at the root or at different levels of the tree. This problem cannot be solved if we randomly select any nod and make it as the root node. Similarly following any random approach will give us the bad results, which may lead to low accuracy.

Hence, in order to solve this problem, many researchers have worked together and found out some solutions. Information gain and Gini index were the two criterion functions suggested by them. These two criterion are used for calculating values for respective or each attribute. When we use the information as a criterion, the attributes are assumed to be categorical and if we consider the Gini index then the attributes are assumed to be continuous.

Input:

I. D is equal to Data partition; it is set of training as it is associated with class labels.
II. Attribute list: It is a set of the candidate attributes.
III. Attribute_selection_method: It is a procedure that partitions the data tuples into the individual classes.

Basic Algorithm (a greedy algorithm):

1. Tree is constructed in a top-down, recursive, divide-and-conquer manner
2. At start, all the training examples are at the root
3. Attributes are categorical (Note that if continuous-valued, they are discretized in advance)
4. Test attributes are selected on the basis of a heuristic or statistical measure

Conditions for stopping partitioning:

- All samples for a given node belong to the same class.
- There are no attributes for further partitioning majority voting is employed for classifying leaf
- There is no sample left.

Pseudo-Code (Decision Tree):

```
if (the sub-table passed to the algorithm is empty)
return NULL;
if (the sub-table passed to the algorithm is homogeneous)
mark this node as a leaf node;
label this node with the value of the last column
else
decide a column to split the table
set the node's split on value to the column name
for all the values that the splitting column can take:
create a new node
```

set the new node as the current node's child node

prune the sub-table

recursively call the function by passing it the new pruned table and the new node

III. Proposed Architecture

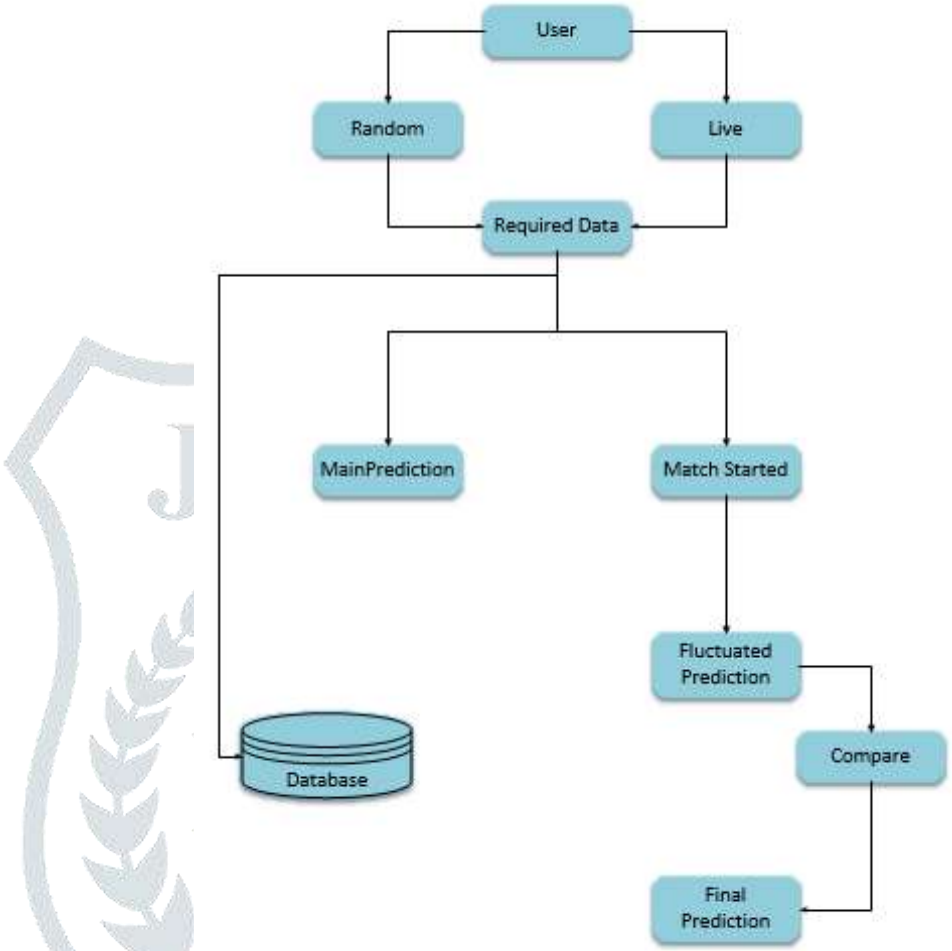

Fig 2: Architecture Diagram

Description in Detail:

1. At the Client side the end user must select the two teams. Depending on the selection of both the teams, the country and the ground where the match will be held will display automatically.

2. After the toss, the playing 11 players of both the teams will be short-listed.

3. Depending on the aspects that are the batting, bowling, and carrier average of the particular player as well as the experience of that player on that ground against rival country. By keeping these averages in consideration, the server will perform the main prediction**.**

4. After the match starts the prediction can differ depending on the following factors:
Batsmen's Performance, Number of wickets, Difference between projected score and the current score.

5. The fluctuated prediction will be compared with the main prediction. (Fluctuated prediction means again the graph will be formed on real time values and will be compared with the graphs that have been made by main predictions.)

6. Whentheresultofthefluctuatedpredictionmatcheswiththeresultofthemain prediction then we have the accurate output.

7. When the result of the fluctuated prediction differs from the result of the main prediction then the system will give a reason of failure

IV. APPLICATIONS

1. Media: The system will work as an expert analyzer

2. Newspaper: It will the descriptive knowledge about the playing 11(players) of the particular match.

3. Online Applications: Example: Dream 11: It will help the users to know the best playing 11 for a particular match and use this knowledge to play on different cricket applications.

V. ACKNOWLEDGMENT

We would like to thank our Guide Mr. S.V. Limkar who helped us throughout the project whenever we stumbled across any obstacle. We would like to also thank the Principal of our institution who gave us the opportunity to work on this project.

VI. REFERENCES

[1] Tianxiang Cui, Jingpeng Li, J. R. Woodward and A. J. Parkes, "An ensemble based Genetic Programming system to predict English football premier league games," *2013 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, Singapore, 2013, pp. 138-143

[2] J. Hucaljuk and A. Rakipović, "Predicting football scores using machine learning techniques," *2011 Proceedings of the 34th International Convention MIPRO*, Opatija, 2011, pp. 1623-1627.

[3] S. Dobravec, "Predicting sports results using latent features: A case study," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2015, pp. 1267-1272.

[4] J. Pan, "Tennis Match Prediction Model Based on Improved D-S Evidence Theory," *2016 International Conference on Robots & Intelligent System (ICRIS)*, Zhangjiajie, 2016, pp. 414-417.

[5] T. L. W. Walls and E. J. Bass, "A regression-based predictive model of student attendance at UVA men's basketball games," *Proceedings of the 2004 IEEE Systemsand Information Engineering Design Symposium, 2004.*, Charlottesville, VA, 2004, pp. 203-208.

[6] D. Miljković, L. Gajić, A. Kovačević and Z. Konjović, "The use of data mining for basketball matches outcomes prediction," *IEEE 8th International Symposium on Intelligent Systems and Informatics*, Subotica, 2010, pp. 309-312.

[7] K. Trawinski, "A fuzzy classification system for prediction of the results of the basketball games," *International Conference on Fuzzy Systems*, Barcelona, 2010, pp. 1-7.

[8] J. Gumm, A. Barrett and G. Hu, "A machine learning strategy for predicting march madness winners," *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Takamatsu, 2015, pp. 1-6.

[9] Y. Saito, M. Kimura and S. Ishizaki, "Real-time prediction to support decision-making in soccer," *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, Lisbon, 2015, pp. 218-225.

[10] B. G. Aslan and M. M. Inceoglu, "A Comparative Study on Neural Network Based Soccer Result Prediction," *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, Rio de Janeiro, 2007, pp. 545-550.