DATA MINING TOOLS IMAPCT ON CRICKET GAMING STRATEGY

*****Dr Ch.Munendar Reddy

MRM Institute Of Management

Abstract

The software and related tools and techniques like data mining, artificial intelligence made everything possible nowadays including born and death predictions of human beings with 90 % accuracy. The medical industry adopted the technology in wide variety to predict Cancer attack chances / decease growth rate etc… The same technology also adopted in sports field especially world popular games like Foot ball and Cricket to analysis particular player strength and psychological ability in the field. The analysis of physiological signals is a hard task that requires the use of specific approaches such as the Knowledge Discovery in Databases process. The application of such process in the domain of medicine has a series of implications and difficulties, especially regarding the application of data mining techniques to data, mainly time series, gathered from medical examinations of patients. Cricket is a sport that contains a lot of statistical data. There is data about batting records, bowling records, individual player records, scorecard of different matches played, etc. This data can be put to proper use to predict the results of games and so this problem has become an interesting problem in today's world. Most of viewers nowadays try to do some sort of prediction at some stage of the tournaments to see which team will eventually win the upcoming matches and thereby the tournament.

Keywords*: Neural Networks, Game Strategy, Data Sets, Algorithms, Machine Learning*

Introduction:

Cricket is being played in many countries all around the world. There are a lot of domestic and international tournaments being held in many countries which play cricket. Cricket is a game played between two teams comprising of 11 players in each team. The result is either a win, loss or a tie. However, sometimes due to bad weather conditions the game is also washed out as

---

*Associate Professor - MBA, MRM Institute of Management, Chinthapallyguda (V), Ibrahimpatan, (M), Ranga.Reddy ..Dist, Telangana 501510

Cricket is a game which cannot be played in rain. Moreover, this game is also extremely unpredictable because at every stage of the game the momentum shifts to one of the teams between the two. A lot of times the result gets decided on the last ball of the match where the game gets really close. Considering all these unpredictable scenarios of this unpredictable game,

there is a huge interest among the spectators to do some prediction either at the start of the game or during the game.

There are different ways to do the prediction. The prediction can be done taking into consideration the player's performance as well as the team performance. There are many unpredictable things that happen in a cricket game like matches being washed out due to rain, a key player getting injured before the game, players changing their teams, etc. Sometimes a key player also gets inured during the game and hence is not able to take further part in the game. All these factors do affect the prediction to some extent. The report discusses a methodology that I followed

for the game result prediction. The methodology consists of first the attribute selection algorithms which trim down the list of attributes to only important ones and then the data mining algorithms which can be applied on those attributes.

Literature:

Luo (2022) in his study adopted a new approach and utilized to collect data on the footwork of badminton players. This study used a deep-learning method to extract two-dimensional (2D) and 3D coordinates of the players' shoes. The model achieved an absolute positioning accuracy of 74%. These data provide valuable insights into the players' movements, which can help improve their performance on the court.

Kulkarni (2021) employed a novel technique to gather data for the classification of different strokes played in table tennis. The authors collected a video dataset of the primary 11 strokes of 14 professional table tennis players and utilized CNN and other machine learning models to classify the strokes. The CNN model achieved an impressive accuracy of 99.37%.

Karmakar (2015) said that , Machine learning models have witnessed a wide adoption in various fields like image processing], text analysis, education], medical data analysis etc., and sports is no exception. As a result, several studies have been presented involving the use of machine learning techniques in sports..

Haghighat (2013) also described the prediction of results in sports using the data mining techniques. But this paper was not specific to any particular sport, rather it was for in general all sports. The attribute selection algorithm that it used was more of an elimination approach where the attributes were eliminated one by one and the classification accuracy is computed. Once a good subset of attributes is achieved, then the eliminated attributes are again added one by one to see if the accuracy improves.

Trawinski (2010) described the prediction of results using a fuzzy classification system. This paper was predicting the results for basketball games. I had used the attribute selection techniques mentioned in this paper for my project. The attribute selection technique proposed in this paper was done using WEKA so it was a good reference point for me too. The wrapper method algorithms and the ranker method algorithms implemented in this paper were also used in my project. But the prediction part was done using the fuzzy classification system and I did not use that system for my prediction part.

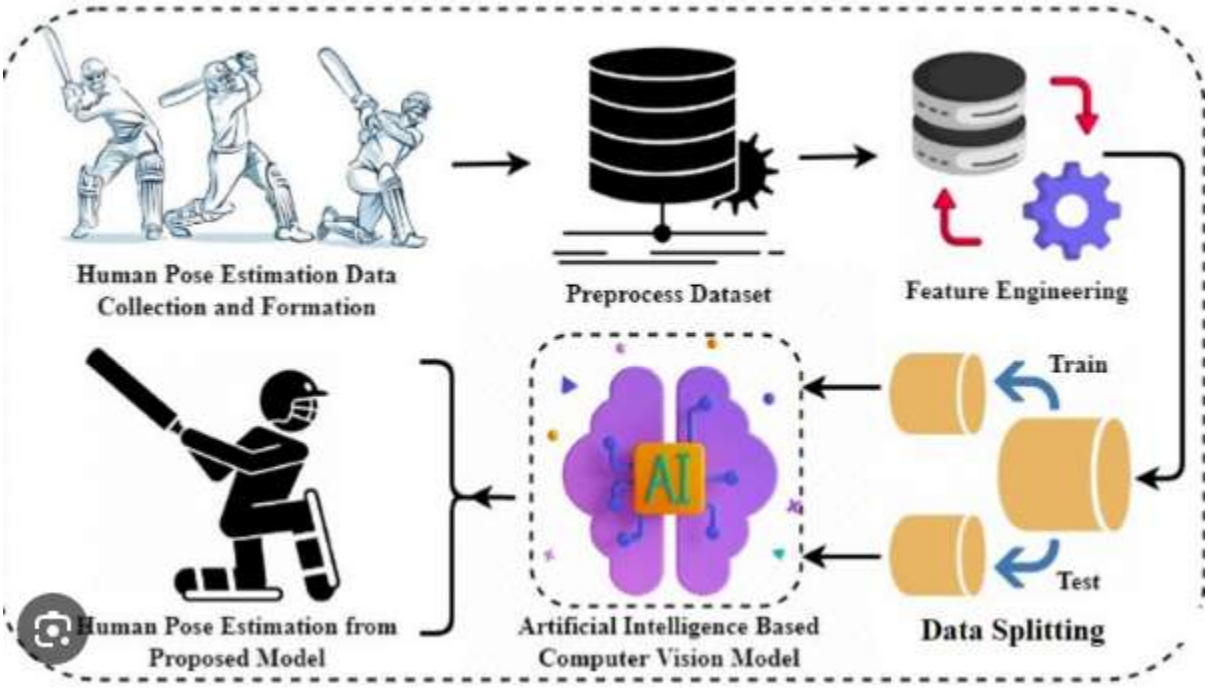
DATA MINING APPROACH IN CRICKET

Data Collection and Extraction.

The Indian Premier League (IPL) is a cricket tournament which is widely famous in our country and has a huge fan following. There have so far been ten seasons of the IPL, with over 500 matches having taken place. The required data set needed for this work has been collected from the internet web source. From this obtained dataset of the IPL matches, only the attributes that are essential to this work needed to be extracted, and so an algorithm to obtain these features alone has been used. The data is then extracted using a Java program which converts the ball-by-ball format into an over-by-over data format.

One of the research makes several significant contributions to the field:

- The study collects a comprehensive video dataset to classify different cricket strokes. In contrast to previous studies that only use image datasets and cover a maximum of five strokes, this study covers eight strokes, including 'flick', 'back foot punch', 'pull', 'cut', 'cover drive', 'straight drive', 'on drive', and 'sweep'.
- A novel technique is employed to extract features from the video dataset. The MediaPipe library extracts seventeen critical points of the human body. Based on these key points, the batsman's stroke is accurately classified.
- The study uses fine-tuned machine learning and deep learning models to classify the strokes based on the extracted feature dataset. Cross-validation is employed to validate the model's performance, ensuring accurate results.



This research provides a more comprehensive and accurate approach to classifying cricket strokes. The novel technique that extracts features from video datasets and utilizes state-of-the-art machine learning and deep learning models helps improve classification accuracy.

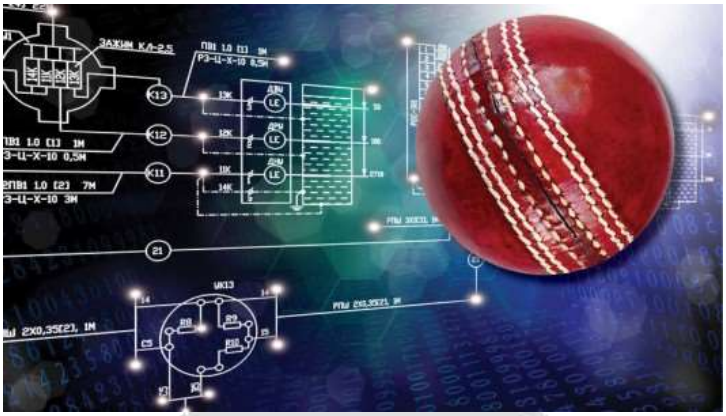
The use of human pose estimation holds several strategic advantages in sports. It can be used by coaches to train players better and enhance their sports performance. Cricket, being the second most popular sport in the world, is liked and followed by billions of people around the globe. Consequently, coaches and players are continuously striving for excellence. The use of machine learning techniques to predict batsmen's strokes can be very influential and useful in this regard. This study collects video data for different strokes and proposes a machine-learning approach for stroke prediction. Different important features are extracted from the preprocessed video data to train machine learning models.

Feature Selection

Feature Selection becomes important when number of features is very high. It refers to selecting the most important features those results in accurate results than the one with all the features. One of the important feature selection methods is Recursive Feature Elimination (RFE). RFE works by recursively building models based on the feature subsets and after the model is built, it removes the feature with very low priority and again builds a model using the remaining feature subsets. The process is continued till all the features are exhausted.

Model Generation

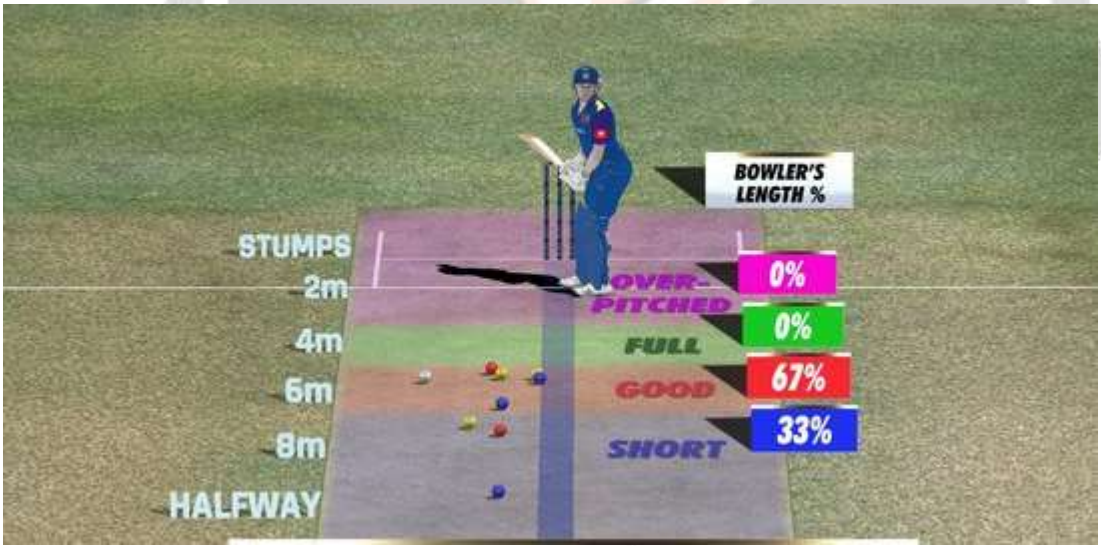
Based on the data obtained, various models required for this research work was generated. The models are generated based n different phases in the match. The different phases in the match are 2-Overs, 5-Overs, 8-Overs, 12-Overs, 16-Overs and20-Overs. These models are used to predict which team will be the winner of the ongoing match. The dataset is loaded first from the corresponding CSV file, and for each phase in the match, a model is create during a different machine learning algorithm, such as Naïve Bayes, SVM, KNN and Random Forest. Then the best model is selected using Cross Validation, and this model is then used to predict the winning team.



Prediction

Once the predictive model is created, the test data is given as input to it and the output is predicted. This is then compared with the ground truth defined in the dataset obtained from the website.

The dataset is loaded from the corresponding CSV file. Apart of the dataset is given as input without feature selection and a part of it is given as input with feature selection to the generated



models. Based on this, different models will be generated based on the different phases of the match, which are2-Overs, 5-Overs, 8-Overs, 12-Overs, 16-Overs and 20-Overs.Based on this, a

prediction on which team wins the ongoing match at that particular phase of the match is obtained.

Data Mining Tools Some of the data mining tools are given below,

1) Artificial Neural Networks (ANN),

2) Rough Set Theory (RST),

3) Statistical Package for the Social Sciences modeler (SPSS),

4) K-means clustering

5) Single Nucleotide Polymorphism (SNP)

Six best open source Data mining tools are given below,

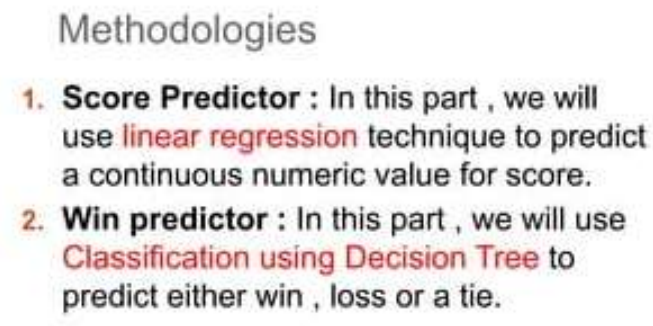
1) Rapid miner 2) Weka 3) R-Programming 4) Orange 5) Knime 6) Natural Languag

Data set 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.



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.

Conclusion:

Our main goals in this paper to develop a model to predict the outcome of an ODI cricket match while the game is in progress. Normally it uses the data of previous matches played between the team in order to design new model. It uses Multiple Variable Linear Regression to design this model. Efficiency and error checking was also done in our work. Using multiple linear regressions, each innings score is predicted at regular intervals and final the winner of the match. This knowledge will help us in the future to design a much more accurate prediction.

References:

- Trawinski,Krzysztof. "A fuzzy classification system for prediction of the results of the basketball games." Fuzzy Systems (FUZZ), 2010 IEEE International Conference on. IEEE, 2010.
- Haghighat, Maral, Hamid Rastegari, and Nasim Nourafza. "A review of data mining techniques for result prediction in sports." Advances in Computer Science: an International Journal2.5 (2013): 7-12.
- Zdravevski, Eftim, and Andrea Kulakov. "System for Prediction of the Winner in a Sports Game." ICT Innovations 2009. Springer Berlin Heidelberg, 2010. 55-63.
- Nadeem, A.; Jalal, A.; Kim, K. Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model. *Multimed. Tools Appl.* 2021, *80*, 21465–21498. [Google Scholar] [CrossRef]
- Song, L.; Yu, G.; Yuan, J.; Liu, Z. Human pose estimation and its application to action recognition: A survey. *J. Vis. Commun. Image Represent.* 2021, *76*, 103055. [Google Scholar] [CrossRef]
- Keshtkar Langaroudi, M.; Yamaghani, M. Sports result prediction based on machine learning and computational intelligence approaches: A survey. *J. Adv. Comput. Eng. Technol.* 2019, *5*, 27–36. [Google Scholar]

- Karmaker, D.; Chowdhury, A.; Miah, M.; Imran, M.; Rahman, M. Cricket shot classification using motion vector. In Proceedings of the 2015 Second International Conference on Computing Technology and Information Management (ICCTIM), Johor, Malaysia, 21–23 April 2015; pp. 125–129.