A Survey on Different Classification technique in Data-Mining

Miss.Mayuri Jadav[1] Prof.Risha Tiwari[2]
[1]P.G. Scholar [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]HGCE, Vehlal

Abstract— In Data mining there are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows: 1. Classification, 2. Prediction. Classification models predict categorical class labels; and prediction models predict continuous valued functions. Here in this paper we have survey different classification method with their advantages and disadvantages. In this paper various classification mechanisms analyzed and their significance given in this survey paper.

Key words: AHP (Analytical Hierarchy Process), Decision tree, IPL player selection, most valuable player

I. INTRODUCTION

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Classification is a data mining technique used to predict group membership for data instances.

Classification by Decision Tree Induction.

1) Decision Tree Induction
2) Attribute Selection Measures
   An attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, D, of class-labeled training tuples into individual classes.
3) Tree Pruning
   When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of overfitting the data.
4) Scalability and Decision Tree Induction
   The efficiency of existing decision tree algorithms, such as ID3, C4.5, and CART, has been well established for relatively small data sets. Efficiency becomes an issue of concern when these algorithms are applied to the mining of very large real-world databases

A. Bayesian Classification

Bayesian classifiers are statistical classifiers. Theycanpredict class membership probabilities, such as the probability that a given tuple belongs to a particular class.

1) Bayes' Theorem
   Bayes' theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century.
2) Naïve Bayesian Classification
   Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case, owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data. Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes' theorem.
3) Bayesian Belief Network
   The naïve Bayesian classifier makes the assumption of class conditional independence, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another.
4) Training Bayesian Belief Networks
   The network topology (or "layout" of nodes and arcs) may be given in advance or inferred from the data. The network variables may be observable or hidden in all or some of the training tuples. The case of hidden data is also referred to as missing values or incomplete data

B. Rule-Based Classification

1) Using IF-THEN Rules for Classification
   The"IF"-part (orleft-handside) of a rule is known as the rule a tecedentor precondition. The "THEN"-part (orright-handside) is the rule consequent.
2) Rule Extraction from a Decision Tree
3) Rule Induction Using a Sequential Covering AlgorithmSequential covering algorithms are the most widely used approach to mining disjunctive sets of classification rules, and form the topic of this subsection.

C. Classification by Backpropagation

1) A Multilayer Feed-Forward Neural Network
   The back propagation algorithm performs learning on a multilayer feed-forward neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.
2) Defining a Network Topology
   Back propagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value.
3) Inside the Black Box: Back propagation and Interpretability

D. Other Classification Methods

1) Genetic Algorithms
   Genetic algorithms attempt to incorporate ideas of natural evolution.
2) Rough Set Approach
   Rough set theory can be used for classification to discover structura lrelationships with in imprecise or noisy data. It applies to discrete-valued attributes. Continuous-valued attributes must therefore be discretized before its use.
3) Fuzzy Set Approaches.

II. LITERATURE SURVEY

A. Prediction of athlete's performance using neural networks: An application in cricket team selection [1].

In this paper authors have proposed a novel approach in which they predict the performance of the all the athletes on the base of their past performance and classify into three categories- performer, moderate and failure. They have used "neural network" method for classify the data. First of all they have taken the dataset and then after they divided the data into bowling and batting. Neural networks in each experiment were trained and tested using primary ratings. After training and testing, each neural network generated its own ratings for all athletes. For giving the ratings for particular cricketer they generate different rules for rating the batsman and bowler. They used for types of the data-set for the classification and then after trained neural network model used for forecast the performance of the each athlete and with the help of the rating generated with the help of the neural network and by applying the heuristic rules they predicted that or recommended that the athlete is eligible or included in the world cup cricket match or not.

The main benefit of this work is that using the neural network and based on the past performance we can predict that how can we make the team as a winning team.

The first drawback is that they recognize that fielding is an important aspect of cricket and they have not included fielding performance of any athlete into their study. Secondly, wicket-keeping is another important aspect of the game and they have not included wicket-keeping analysis in their research.

B. Forecasting test cricket match outcomes in play [2].

In this paper authors have proposed a method in which they forecasts match outcomes in test cricket in play, session by session. Match outcomes probabilities at the start of the session which are forecast using the method of the classification is multinomial logistic regression models. The probabilities are defined by different criteria like certain aggressive or defensive batting strategy for the coming session. For that they have to noticed past matches in-play effects like strengths of teams, a ground effect, home field advantage, outcome of the toss, score or lead, overs-used, overs-remaining, run-rate, and wicket resources used and after using that they derives some probabilities outcomes like win, draw, or loss. The outcomes describes that the lead has a small effect on the match outcome early on but is dominant later; pre-match team strengths, ground effect and home field advantage are important predictors of a win early on; and wicket resources used remains important throughout a match. The aim of this work is to provide a quantitative means of forecasting match outcomes in play.

The benefits of this work is for the start of the session positions has two advantages Firstly, some progress can be made with the quantitative analysis of the problem. Secondly, the models can guide the team captains and management with respect to batting and bowling strategies in each session.

One of the drawbacks of their work is that if both teams lose wickets in the first day's play then their models are not able to take the number of wickets down into account for the team batting second. Data on more matches would helpful for overcoming this problem.

C. Valuing Cricketers Using Hedonic Price Models [3].

In this paper authors have proposed an approach for find most valuable athlete by not on the base of the their performance that how many runs they had made or how many wickets they had taken but also they add additional information like that he is young or old because obviously young athlete has more stamina power compare to the old athlete and also one attribute is price means which athlete got the higher price in the auction process so based on this criteria they decided which athlete is strong for select in the match. They have used regression model for prediction based on the past data they predict the new values. They have included age of the athlete as an additional regressor in the base-level regression:

PRICEi = a0 + a1 *RUNSRi+ a2 *WKTSRi + a3 *AGEi + ui

Where,

PRICEi = winning bid for a athlete (measured in US$).

RUNSRi = ratio of runs scored by a athlete in 1-day and T20 formats to the total runs scored by all the 75 athletes, expressed as a percentage.

WKTSRi = ratio of wickets taken by a athlete in 1-day and T20 formats to the total wickets taken by all the 75 athletes, expressed as a percentage.

AGEi = age of the athlete (measured in years).

Based on this probability equation they decided which athlete will be selected for the next match on the basis of the total value of the price.

The benefit of this work is that we can get the best athlete and he is not forecast only base on the performance but we get the idea that which athlete is famous in recent and also we can get idea that which athlete is younger with high efficiency.

The drawback of this work is that as we know that winning team had the highest rate of return. They have not been able to establish whether the IPL bidding process is affected by the winner's curse. The negative rates of return for some of the franchises suggest that the winner's curse might well be a problem.

D. A Hedonic Model of Athlete Wage Determination from the Indian Premier League Auction: Further Evidence [4].

In this paper authors have proposed the approach in which they built the system in which they determine the best athletes for the application to cricket's Indian Premier League (IPL). They distinguish between personal characteristic and playing ability factors, and with respect to the former, between ability in different forms of the sport. They find a number of interpretable variables that have explanatory power over auction values, while decomposition according to batting and bowling specializations produces very different results depending on the use of either Test or One-Day International (ODI) variables. They calculated pair-wise correlation coefficients between the possible pairs of independent variables. Initially, two models were estimated for all 80 athletes purchased in the auction – one using Test (and Domestic) career statistics and the other using ODI statistics, as it is not clear *a priori* which set of statistics contain more information about the athlete's ability to perform in the IPL and hence which set bidders will give greater weight to. Since the volume of Twenty20 matches played up to February 2008 is thin, it is inadvisable to use career statistics in that form of the game, and rather we use simply the number of games

played, as any experience at Twenty20 would be considered potentially valuable.

The benefits of this work is that they interpret the presence of serial correlation as some evidence of overbidding for star athletes and underbidding for lesser athletes. Secondly, ODI statistics seem to provide more informational content about batsmen than Test statistics, while the inverse is true for bowlers.

The drawback of this system that both one day match (ODI) and test cricket match both gives different result so sometimes it may be produced confusion and we can't forecast the result correctly based on these parameters appropriately.

E. Quantifying individual performance in Cricket – A network analysis of Batsmen and Bowlers [5].

In this paper authors said that quantifying individual performance in the cricket is always depended on the runs scored by batsmen and wickets taken by bowlers. Traditionally the batsmen and bowlers are rated on their batting or bowling average respectively. However in a game like Cricket it is always important the manner in which one scores the runs or claims a wicket. Scoring runs against a strong bowling line-up or delivering a brilliant performance against a team with strong batting line-up deserves more credit. An athlete's average is not able to capture this aspect of the game. So they have presented a refined method to quantify the 'quality' of runs scored by a batsman or wickets taken by a bowler. They explore the application of Social Network Analysis (SNA) to rate the athletes in a team performance. They generate directed and weighted network of batsmen-bowlers using the athlete-verses-athlete information available for Test cricket and ODI cricket. For that they take the datasets in which the data of athlete-verses-athlete contains the information of runs scored by a batsman against every bowler he faced and how many times he was dismissed by the bowlers he faced. Thus higher batting average reflects higher 'quality' of a batsman. Similarly, bowling average is defined as the number of runs given by the bowler divided by the number of wickets claimed by him. Thus lower bowling average indicates higher ability of the bowler. This information is used to generate the network of interaction among bowlers and batsmen in cricket matches. They represent a performance index of a batsman (PIB) against a bowler given by the following equation

$$PIB = A_{Ba} / C_{Bo}$$

Where A_{Ba} is the batting average of the batsman against the bowler he faced and C_{Bo} refers to the career bowling average of the bowler.

The advantage of network analysis is that it doesn't introduce these 'constraints' and yet provides consistent results.

The drawback is they don't consider the fielding abilities or wicket-keeping abilities of the fielders. Nevertheless a network based approach could address the issue of relative performance of one athlete against other.

F. AHP-Neural Network Based Athlete Price Estimation in IPL [6].

In this paper authors have proposed an approach to find the best athlete for choose the most valuable athlete based on Analytical hierarchy Process (AHP) and Artificial Neural Network (ANN) for estimation of athlete price in IPL.Based on expert view several key features are chosen for cricket athlete price calculation in IPL twenty-20 cricket tournament. Initial weights of attributes are calculated through AHP. Back propagation neural network trains a pre normalized performance dataset of last three years IPL statistical dataset of 226 athletes. Finally, our proposed methodology gives a systematic way to select the important attributes and calculate the weights based on expert opinion to measure the optimal price for a athlete which will help the IPL team owner to select the athlete according their budget and strategies. The athlete's price estimation depend on three basic features namely as-

- Athlete's Performance Appraisal
- Athlete's Experience Contribution
- Athlete's recent form

The benefit of this work is this model helps us to handle the complexity and selecting the attributes for athlete price calculation.

The drawback of this work the selection of a winning cricket team with athletes selected from optimally found criteria values and minimum budget is not possible in this approach.

G. Multi-Criteria Decision Tree Approach to Classify All-Rounder in Indian Premier League [7].

In this paper authors proposed the system in which multi-criteria analysis plays a vital role to measure the performance of cricketers and Decision tree technique helps us to classify in very efficient manner. This research makes use of technique for order preference by Similarity to Ideal Solution (TOPSIS) method to produce the overall performance of the all-rounder of Indian Premier League (IPL) T-20 session-III cricket tournament. The result of TOPSIS method is then used to classify all-rounder in four different categories by using Decision tree. Finally, they have a proposed a multi-criteria decision tree approach which provides accurate and efficient data classification upon the athlete's performance. Here we have defined classification tree which is used in their approach.

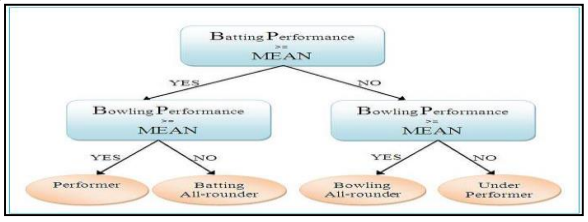

Fig. 1: The prediction tree for classifying the all-rounder [7].

The benefit of this work is that the role of the all-rounder in twenty-20 cricket is much more than other category athletes for team's better performance and this approach is used to find best all-rounder among the entire athlete. The drawback of this system that it is work for the find best all-rounder athlete but it is not work for best fielder or particular best bowler or batsman.

H. Statistical Based Multi-Criteria Decision Making Analysis for Performance Measurement of Batsmen in Indian Premier League [8].

In this paper authors represents an approach in which their aims to measure the performances of batsmen during first three session of Indian Premier League (IPL) Twenty-20 cricket tournament. Technique for Order Preference by

Similarity to Ideal Solution (TOPSIS) method for multi-criteria decision making has been used to evaluate the performances of batsmen. Then analytical Hierarchy Process (AHP) has been used for calculating the weights of the batsmen's criterion and One-way Analysis Of Variance (ANOVA) has been used to measure the contribution of the different criterion which is then combined with the AHP results to get the modified weight of each criterion. Finally, this work proposed a statistical based multi-criteria decision making analysis which provides a comparison between the batsmen in three IPL and evaluate the overall performances of batsmen. Here the batsmen of IPL are the selective alternatives and the following conditions are used for selection of athletes.-

- Batsmen who played all three session (I, II, III) of IPL.
- Batsmen played at least three innings in a particular IPL session.
- Batsmen who scored at least 10 runs in a particular session.

The benefit of this work is that for fare selection of athletes this proposed performance measurement of batsmen plays a very important role. Their proposed methodology is also useful in IPL athlete's auction to bid the appropriate value of the athlete according their performances. The drawback of this work is that this system is only used for the find best batsman but this approach is fail to find best bowler or all-rounder or fielder.

I. Players Valuation in Indian Premier League Auction using Data Mining Technique [9].

In this paper authors represents some predictive models for guessing or predicting the selection of an athlete in the Indian Premier League, a cricket league, based on each athlete's past performance. Using One-Day International (ODI) variables and T-20 variables of both batting and bowling, they have found a number of interpretable variables that have explanatory power over auction values. The models that are developed can help decision makers during the auction to set salaries for the athletes. They have created an individual dataset for bowler, batsman and all-rounder. The batsman dataset consists of 40 attributes, the bowler dataset consists of 32 attributes and the all-rounder dataset consists of 64 attributes. Data mining tools are used to predict the base price group for athletes which will prove to be automated and a fair way of doing it. The athlete's past performance has been used to predict their base price. Three algorithms are applied namely decision tree, naïve Bayes and MLP on the data set to predict the base price of athletes. The algorithms are applied in order on batsmen, bowlers and all-rounder data set respectively. These models have the ability to build a talented team with minimum cost.

The main advantage of this work that this approach is cost efficient to find the base value of the athlete. In this research three different classification methods are used so it will also help in which method is accurate for the prediction. The drawback of this system is that here in this work they have proposed only classification approach but from that we will not predict the which athletes performs together very well or which athlete is best here we get best group only and best fielder is also not defined in this approach.

	RESEARCH PAPER	BENEFITS	DRAWBACKS
1	Prediction of athletes performance using neural networks: [1]	- uses neural network - based on the past performance predicts winning team.	- Fielding, wicket-keeping is not included.
2	Forecasting test cricket match outcomes in play [2]	- Work is for the start of the session positions uses quantitative analysis of the problem.	- if both teams are losing wickets the algorithm takes time -Data is limited.
3	Valuing Cricketers Using Hedonic Price Models [3]	- Forecast based on performance but -high efficiency.	- Winning team had the highest rate of return. -winner's curse is problem.
4	A Hedonic Model of Athlete Wage Determination from the Indian Premier League Auction: [4]	- Interpret evidence of overbidding and underbidding for star or lesser athletes. -Informational content about batsmen inverse is true for bowlers.	- ODI and test cricket match both gives different result so forecast I not proper.
5	Quantifying individual performance in Cricket [5]	-Network analysis is that it doesn't introduce these 'constraints' and yet provides consistent results.	- fielding abilities or wicket-keeping abilities of the fielders are not taken.
6	AHP-Neural Network Based Athlete Price Estimation in IPL [6]	- handle the complexity and selecting the attributes for athlete price calculation.	-Minimum budget is not possible in this approach.
7	Multi-Criteria Decision Tree Approach to Classify All-Rounder in Indian Premier League [7]	- all rounder in twenty-20 cricket is much more than other category athletes for team's better performance	- Best all-rounder athlete but it is not work for best fielder or particular best bowler or batsman.

8	Statistical Based Multi-Criteria Decision Making Analysis for Performance Measurement of Batsmen in Indian Premier League [8]	- The benefit of this work is that for fare selection of athletes this proposed performance measurement of batsmen plays a very important role.	- This system is only used for the find best batsman but this approach is fail to find best bowler or all-rounder or fielder.
9	Players Valuation in Indian Premier League Auction using Data Mining Technique [9]	- Cost efficient to find the base value of the athlete. -Research three different classification methods for accurate prediction.	- Proposed only classification approach will not predict the which athletes performs together very well best fielder is also not defined in this approach.

III. PROPOSED WORK FRAMEWORK AND METHODOLOGY

A. *Flowchart of Proposed methods.*

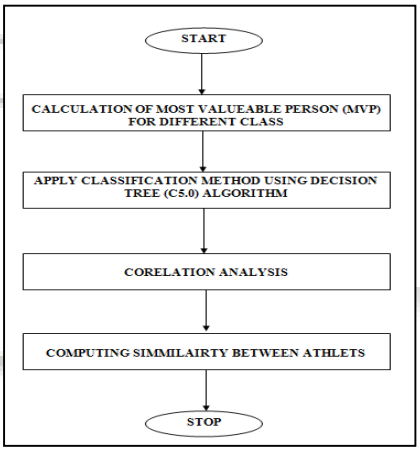

Fig. 2: Flow of proposed methodology framework

Proposed Algorithm for players classification

Step 1: Load database

Step 2: According to the requirement calculate MVP (Most Valuable Player)

Calculation of MVP formula is TRP = PBT + PBW + PEX

Calculate PBT formula is

PBT = (Batting Average *0.3) + (Batting Strike-Rate * 0.4) + (floor (Number of Hundreds) * 0.1) + (Number of Fifties * 0.2)) / 10)

If that the bowler must have bowled minimum100 bowls in his IPL career then;

Calculate PWB formula is

PBW = (((300 / Bowling Average) + (200 / BowlingStrikeRate) + (300 / Economy) + (floor (Numberof4 - wicketshaul) * 0.1 + (floor (Numberof5 - wicketshaul) * 0.1) / 10)

Calculate PEX formula is

PEX = (Number of Matches Played / Total Number of Matches in IPL so far)

Calculation of MVP calculation

If PBW = 0 then; MVP= (8 * PBT * (BARP) + (PBW * BARP) +PEX * ERP) / (TRP *10)

IF PBT/PBW >= 2 THEN MVP

= (7 * PBT * (BARP) + (2 * PBW * BORP) + (PEX * ERP)) / (TRP * 10)

IF PBW/PBT >= 2 THEN MVP

= (2 * (PBT * BART) + 7 * PBW * BORP + (PEX * ERP))/ (TRP / 10)

Otherwise,

MVP = (9 * PBT * (BARP) + (9 * PBW * BORP) + (2 * PEX * ERP)) / (TRP * 20)

Step3: Classify players classification method using decision tree (c5.0) algorithm

Step 4: Lift for each player = LF (P1, P2)

$$\text{lift (P1, P2)} = \frac{P(P1U\ P2)}{P(P1).P(P2)}$$

lift (P1, P2) < 1, negatively correlated

lift (P1, P2) >1, positively correlated lift (P1, P2) = 1, P1 is independent of the occurrence of P2 and there exists no correlation.

Step 5: calculating Similarity and Dissimilarity of two players

Formula of similarity calculation

$$\text{d (P1, P2)} = \frac{b+c}{a+b+c+d}$$

Formula of Dissimilarity calculation

Similarity (P1,P2) = 1 - d(P1,P2)

Where, a =00, b=01, c=10, d=11

Procedure CLASSIFICATION (PBT, PBW)

5.1 if (PBW=0) or (PBT/PBW)>= 4) then

PBT>=7.0 then select A as the batsman

PBT>6.0 then select B as the batsman

PBT>5.0 then select C as the batsman

Then Batsman D

5.2if (PBW/PBT)>=1.25 then

PBW >=7.0 then select as Bowler A

PBW >6.0 then select Bowler B

PBW >5.0 then select Bowler C

Then Bowler D

else

5.3 if (PBT+PBW/2)>=6.5 then select All Rounder A

(PBT+PBW/2)>=6.0 then select All Rounder B

(PBT+PBW/2)>=5.5 then All Rounder C

All Rounder D

Step 6: Stop

IV. RESULT ANALYSIS

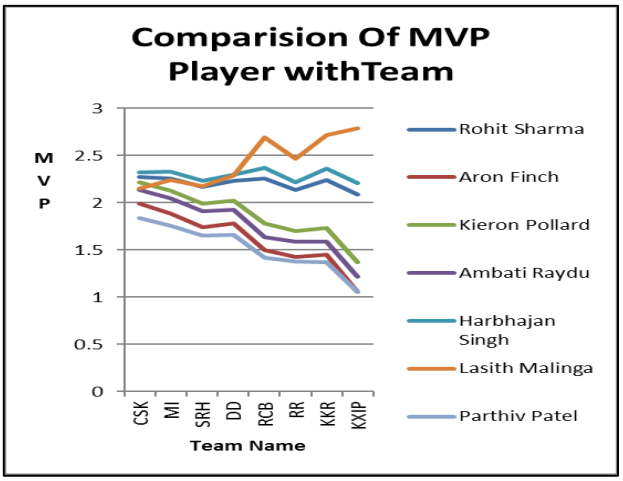

Fig. 3: Comparison of MVP Players with team

Above figure 3 shows the MVP value for each player considering them in different teams according to requirements

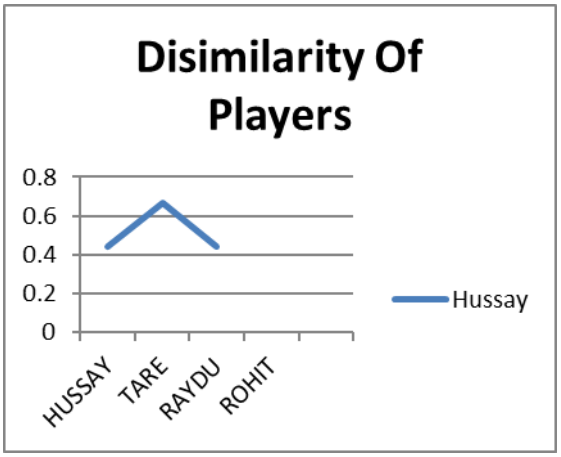

Fig. 4: Dissimilarity of players

Above figure 4 shows the Dissimilarity value for each player with other

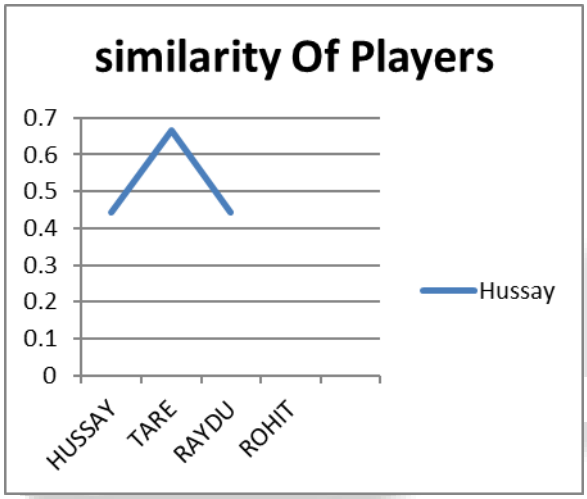

Fig. 5: Similarity of players

Above figure 5 shows the Similarity value for each player with other.

A. *MVP calculation for Athletes*

In the implementation part we can see that we have different teams and we know that all team has different requirements for bowler, batsman and all rounder according to their requirements for player. Suppose Rohit Sharma is selected as batsman for team MI then the MVP calculation is 2.26 and when we select same player and take different team than the MVP calculation is different. So in this way we can say that by our novel approach we can find most valuable player for different team dynamically in IPL.

B. *Correlation analysis using lift for Athletes*

In this part we can find which player is related to the teams means if suppose Rohit Sharma is a team member of MI then here we can check that this player is negatively correlated or positively correlated or independent so if we don't take this player then we can check the affect of the player that if we don't take this player then the team will be in loss or not so by using this work we can find easily which player is most important for the team and which player is not important for the team.

C. *Similarity Measurement:-*

Similarity measure as the name suggests gives the idea about the common features of cricket among two players. We have used equation from "data mining stream cluster analysis" which is to find similarity and dissimilarity between two players of asymmetrical binary output. As we know that IPL is totally auction based procedure. So suppose an owner Select's Rohit Sharma first and another owner wants to select the same player then which is not allowed but considering the similarity measurement player similar to Rohit Sharma can b optedas an option in the team. So this is also very important parameter to choose player in IPL.

V. Conclusion

This paper has surveyed the literatures on IPL player selection models across diverse disciplines. The Disadvantage of each of the protocol has been pointed out. We have attempted to integrate our understanding across the surveyed literatures any tried to find out the one system proving the IPL player selection.

References

[1] S. Akhtar and P. Scarf, "Forecasting test cricket match outcomes in play," International Journal of Forecasting, vol. 28, no. 3, pp. 632–643, Jul. 2012.
[2] S. R. Iyer and R. Sharda, "Prediction of athletes performance using neural networks: An application in cricket team selection," Expert Systems with Applications, vol. 36, no. 3, pp. 5510–5522, Apr. 2009.
[3] A. Karnik, "Valuing Cricketers using Hedonic price models," Journal of Sports Economics, vol. 11, no. 4, pp. 456–469, Nov. 2009.
[4] L. J. A. Lenten, W. Geerling, and L. Kónya, "A hedonic model of player wage determination from the Indian premier league auction: Further evidence," Sport Management Review, vol. 15, no. 1, pp. 60–71, Feb. 2012.
[5] S. Mukherjee, "Quantifying individual performance in cricket — A network analysis of batsmen and bowlers," Physica A: Statistical Mechanics and its Applications, vol. 393, pp. 624–637, Jan. 2014.
[6] P. K. Dey, A. Banerjee, D. N. Ghosh, and A. C. Mondal, "AHP-Neural network based player price estimation in IPL," International Journal of Hybrid Information Technology, vol. 7, no. 3, pp. 15–24, May 2014.
[7] P. K. Dey, D. N. Ghosh, and A. C. Mondal, "IPL team performance analysis: A multi-criteria group decision approach in fuzzy environment," International Journal of Information Technology and Computer Science, vol. 7, no. 8, pp. 8–15, Jul. 2015.
[8] Pabitra Kumar Dey,DipendraNath Ghosh,"Statistical Based Multi-Criteria Decision Making Analysis for PerformanceMeasurement of Batsmen in Indian Premier League", International Journal of Advanced Research in Computer Science, Volume 3, No. 4, July- August 2012.
[9] Prince Kansal, Pankaj Kumar, Himanshu Arya, Aditya Methaila, "Player Valuation in Indian Premier League Auction using Data Mining Technique", 2014 IEEE.
[10] http://stats.espncricinfo.com/ci/engine/stats/index.html

