International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 06, June 2017
www.ijates.com
ijates
ISSN 2348 - 7550

AN APPLICATION OF MULTIRELATIONAL CLUSTERING WITH DOMAIN EXPERT GUIDENCE FOR VERIFYING SESSION FIXING IN CRICKET

K.Ananthapadmanabha[1], Dr.K.Udayakumar[2]

[1]Research Scholar, PP COMP.SCI.ENG.0251, Computer Science & Engineering
Rayalaseema University, Kurnool,(India)
[2]Principal and Professor, Department of Computer Science and Engineering
Adarsha Institute of Technology, Bangalore, (India)

ABSTRACT

According to Kevin Peterson a well known international repute sport's advocate and editor of lawinsport.com, the greatest threat for sport world wide is match fixing. Different types of match fixing are spot fixing, session fixing and bracket fixing. Session fixing is a type of match fixing in which the match is divided into different sessions and in each session, betting opportunities are made available for general public. To verify session fixing in cricket latest technology like sports data mining can be used. Sports Data Mining deals with sports data in all domains of sports like football, cricket, volleyball, hockey and expertise available in related sports domain helps in analyzing sports data. Session fixing is done by dividing each inning of the cricket match into different sessions like a T20 cricket match inning can be divided into four sessions each of five over's and number of session in a match of two inning will be eight sessions. To check session fixing in this paper we are applying Multi relational clustering with Domain expert guidance.

Keywords : Cricket, Domain expert guidance, Multi relational clustering, Session fixing, Sports Data Mining.

I. INTRODUCTION

The association that exists between different players, umpires and officials participating in a cricket match can be represented in different relations. Multi relational clustering is applied on these relations to partition these relations into different clusters. Domain experts like former captain or senior commentators have good knowledge about problems like session fixing, what inside information is needed for performing session fixing etc. This knowledge can help in analyzing whether any match is session fixed or not.

II. SPORTS DATA MINING

In today's world, sports is not only played for entertainment, It has moved beyond entertainment and it is a multi trillion dollar industry. With many sports based enterprises investing multiple billions of dollars for their operations. This industry today has huge volumes of sports data across all domains of sports. This data can be with respect to individual player performance, team performance, tournament details and game details. All these sports data can be used for professional purposes like team selection, captain decision making process, for coaching or managerial decision making process. It can also be used for trend analysis, sports management,

talent recognition, Sports sponsorship, prediction of match result outcomes, analyzing controversies in sports like doping scandals, match fixing scandals to name a few problems. It also helps in fine tuning fitness level of players and enhances teams performance. In any team decision making process this sports data can be utilized for competitive advantages to be ten steps ahead of their opponents. It assists them in designing match strategies, team selection, analysis of opponent teams strength and weakness. With each sport there is huge amount of expertise available. This helps in analyzing sports data and contribute for development of research and development in this area. This sports data have hidden relationship which when mined provides competitive advantages.

Sports data may be in the form of comments and reviews stored in social media like Twitter, Face book to name a few which can be analyzed using data mining techniques like Opinion mining or Sentiment analysis, to understand background knowledge using Ontology based mining techniques. Its application includes Ontology mapping, expertise matching, Opinion spam detection etc. Review mining can be conducted for verifying reviews stored in sports reviews on the social media by both viewers and domain experts on a particular sports based on event, topic or game.

Sports Data Mining consists of different tools and techniques to measure individual player performance and team performance. It has given an opportunity to automate sports data from human level to automated Information retrieval and storage system for extracting knowledge from sports data. Sports data analysis requires new and novel techniques which are completely different from classical data mining procedures and techniques.

III. LITERATURE SURVEY

Not much work is done on session fixing in cricket. In a paper titled " Match fixing network analysis for verifying nearness among internal participants of a cricket match" authored by the same authors in IEEE proceeding they have highlighted the role of internal participants of a cricket match and significance of match fixing network for conducting match fixing. Here they have proposed three algorithms for defining nearness based on geographic proximity. These algorithms are ApartmentNN() Algorithm, OfficeNN() and TransportationbasedNN(). In another paper authored by the same authors they have focused on role of Iceberg diagrams in verifying match fixing in cricket.

IV. METHODOLOGY

In a game of cricket there will be N number of players and umpires participating out of which K number of clusters can be created. But not all K number of clusters are match fixers or outliers. Between players, umpires, officials, bookies and gamblers there will be some form of association .This association can be represented in multiple relations. Also multi relational clustering process is used to partition these different data objects into a set of clusters. Here user guidance in clustering and tuple ID propagation is used to avoid physical joins.

V. N DIMENSIONAL SPACE

N Dimensional Space is a space in clustering and nearest neighbor used to define what is near and what is far away based on distance calculated. This distance is Euclidean distance given by the formula

$$D_{(x,y)} = \Sigma ((xi - yj)2)1/2 \quad \ldots. (1)$$

Real world problems like match fixing in cricket consists of N dimensions where each predictor that is used can be considered to be a new dimension .

VI. MULTI RELATIONAL CLUSTERING WITH DOMAIN EXPERT GUIDANCE

One major problem in multi relational clustering is there are too many predictors or attributes in different multiple relations. Also a specific set of these attributes are relevant to a specific clustering task. End users with their application experience have a specific event or pattern in mind using which they would like to find a pattern. Also they have a good idea about application requirements and data semantics. They know which attributes are relevant and which attributes are irrelevant. To empower end users to conduct their own prediction there is a need to include multi relational clustering with domain expert guidance . Here domain expert guidance will be in the form of simple queries which is used to improve the efficiency and quality of high dimensional multi relational clustering.

VII. N DIMENSIONAL SPACE REPRESENTATION.

A multidimensional representation of the data together with all aggregates is known as Data Cube. A Data Cube may have either more or fewer than three dimensions. It is a generalization of Cross Tabulation. Multidimensional data analysis consist of viewing the data as a multidimensional array and aggregating data for better analysis of structure of data. Multidimensional data analysis supporting relational databases are ROLAP Systems. There are other types of multidimensional data analysis like MOLAP Systems.

VIII. SIMPSON'S PARADOX

In multidimensional clustering there are too many attributes in different relations. It is important to exercise caution when interpreting the association between attributes because the observed relationship may be influenced by the presence of other confounding factors like hidden variables that are not included in the analysis. These hidden variables may cause the observed relationship between a pair of variables to disappear or reverse its direction. This phenomenon is called Simpson's paradox. It leads to generation of spurious tuples. To avoid the generation of such spurious patterns proper stratification is required.

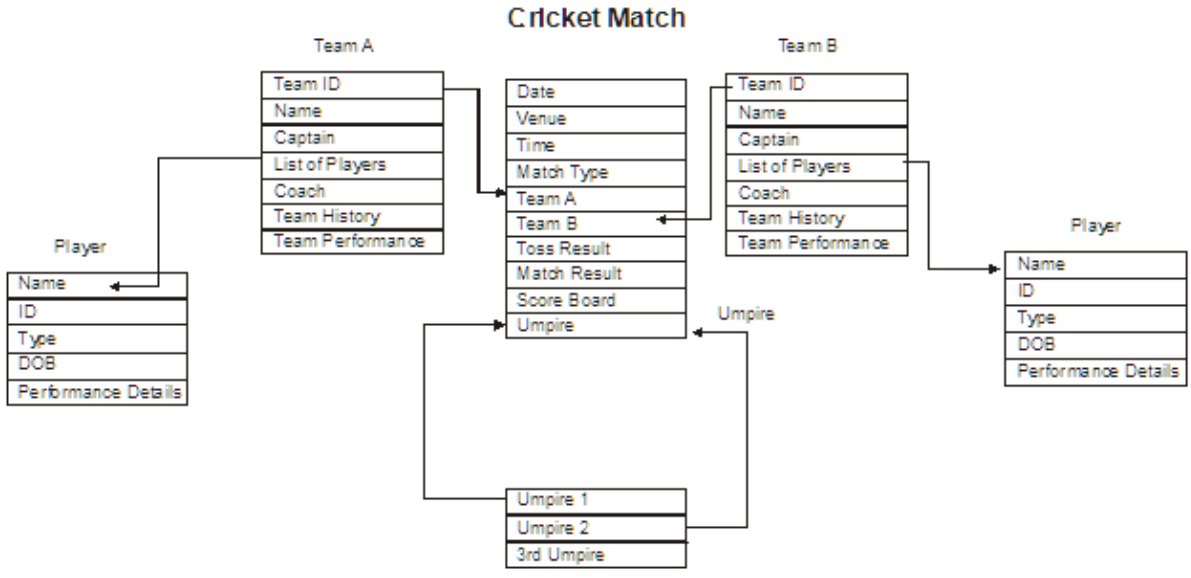


Fig 1: Multi Relational Schema of a Cricket Match

With respect to Match Fixing in cricket consider team A. In order to cluster team A players, attributes like Matchtype(Test, ODI and T20) participated by players, performance of players, coach details are considered.

A user is interested in clustering players based on certain aspect of information like clustering based on T20. Here users have a clear picture about their application requirement and data semantics. The user who provides guidance may be a experienced player or an expert commentator. His guidance in the form of simple query is used. Now consider user queries with a target relation with one or more pertinent attributes to specify goal of the user. A multi relational attribute A is defined by a join path An attribute R.A of R and an optional aggregate operator like Max, Min, Count, Average. Multi relational attribute A is represented by A.joinpath.A.attr.Aggregater in which A.attr.Aggregater is optional A multi relational clustering can be a categorical feature or a numerical one. If A is a categorical feature then for target tuple t, t.A represents the distribution of values among the tuples in R that are joinable with t. If As a numerical, then it has a certain aggregation operator like max, min, count, average and t.A is the aggregated value of tuples in R that are joinable with t. In Multi relational clustering process, search for pertinent attribute across multiple relations.

Major challenges in the search process are

1. The target relation R can join with non target relation R via many different join paths and each attribute in R can be used as a Multi Relation Attribute. But it is not feasible to conduct exhaustive search.
2. Among the huge number of attributes, some are pertinent to the user query where as many other attributes are irrelevant. So there is a need to identify pertinent attributes while avoiding aimless search in irrelevant regions in the attribute space.

Steps taken to overcome challenges of multi relational clustering are

1. To confine the search process. Consider the relational schema as a graph with relations as nodes and joins as edges of the graph. Apply heuristic approach of search which starts search from the user specified attribute and the repeatedly search for useful attribute in the neighborhood of existing attribute.
2. To identify neighboring attributes as pertinent attribute check how attribute cluster target tuples. The Pertinent attribute are selected based on their relationship to the user specified attribute. If two attributes cluster tuples vary differently, then the similarity is low and they are not related. But if two attributes cluster tuples very similarly their similarity is high and they are considered related. If two attributes cluster tuples in almost the same way their similarity is very high and they may represent redundant information.
3. From the set of pertinent features found select a set of non redundant feature so that similarity between any two features is not greater than a specified Maximum.
4. For evaluating the similarity between attributes the data structure used is Similarity vector which is defined as follows

Suppose there are N target tuples t1.t2…..tn Va be the similarity vector of attribute A. It is an N dimension vector that indicates the similarity between each pair of target tuple ti and tj based on A.

IX. SESSION FIXING

In an ODI cricket match, a team's inning can be divided into three sessions. The first two sessions are of fifteen over's while the third session will be of twenty over's. In a two inning ODI match there will be six sessions. For each of these sessions, bets are accepted by the bookies based on permutation and combination of few important critical information . This information corresponds to teams game plan, team composition, pitch report, weather

report etc. Such information is called Inside Information. Bookies offer odds on session betting. So there are more betting options for bookies when compared to match fixing where they can provide betting option only on the end result of the match. Inside information is extracted from teams dressing room and this information helps bookies in understanding its impact on a particular session of the match or the final end result of the match.

X. ALGORITHM SESSIONFIXING()

//Purpose : To identify occurrence of session fixing in a cricket match

//Input : Inside information from dressing room and score board details

//Output : To verify any session fixing is done or not

Step 1 : Decide in advance players from which team who are part of dressing room who will pass information to the bookies.

Step 2 : Divide the T20 cricket match into sessions. Check the inside information to decide its impact on every session of the match

Step 3: For each session in the match

Step 4: Decide signal of underperformance by player of a specific team like displaying a towel on the trouser of the player.

Step 5 : Decide betting rates based on inside information.

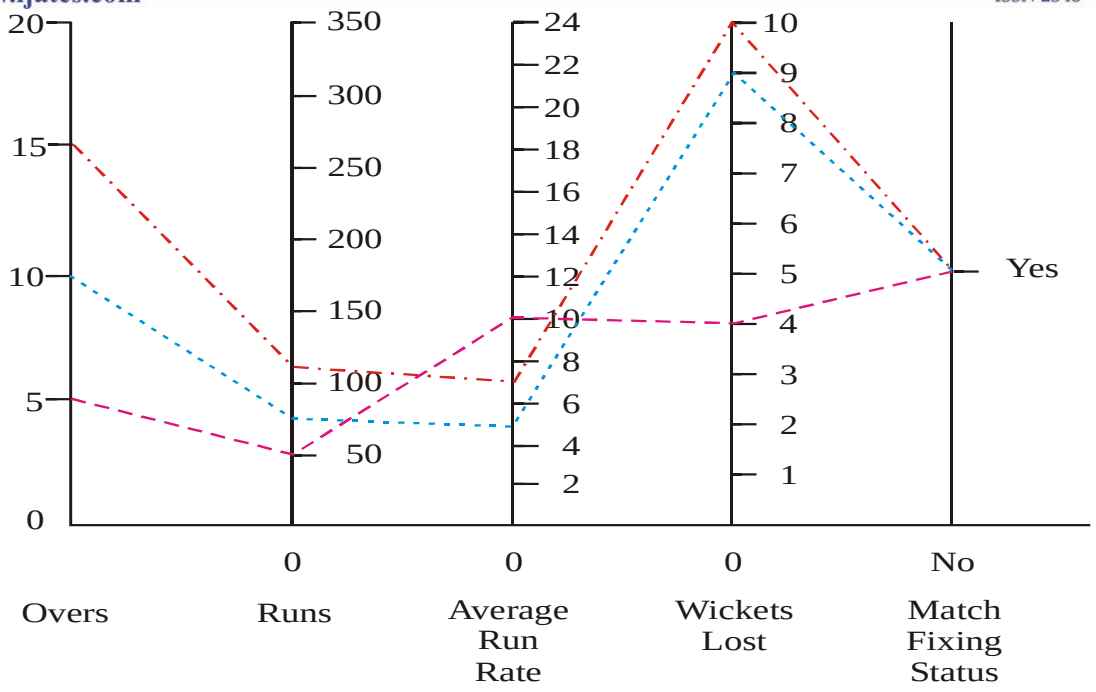
Step 6 : If the specified number of runs in the session and number of wickets to be lost is decided between the bookies and players and if the players are ready for underperformance players have to give signals to the bookies as decided earlier.

Step 7 : Players to underperform their role as decided earlier.

Step 8 : If players do not give signals to bookies even though they are ready to underperform the bookies may not involve in betting for that session.

XI. RESULTS

In a cricket match there will be two teams namely team A and team B. If team A wins the toss and elects to bat first then team A is called target setter and team B is called target chaser. For each team there is a need to draw Iceberg diagram which is based on parallel coordinate system and can include any number of dimensions. Here dimensions included for target setter team are overs, runs scored, average run rate, wickets lost and status of match fixing. For target chaser team all the above dimensions are included along with an additional dimension called target run rate.



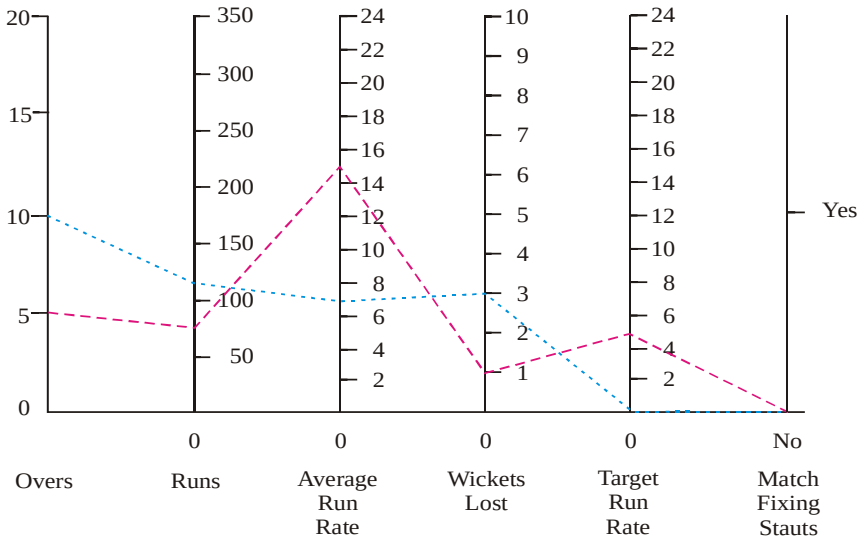
Graph 1: Iceberg Diagram for Team A (Target Setter) in Match 1

Graph 1 represents Iceberg diagram for team A as target setter in match 1 between team A and team B. Team A in its first 5 overs (overs 1-5) of its batting innings scores 50 runs by loosing 4 wickets at an average run rate of 10 runs per over.

In the next five overs(overs 6-10), team A scores 25 runs by loosing 5 wicket at an average run rate of 5 runs per over.

In the next five overs (overs 11-15), team A scores 35 runs by loosing 1 wicket at an average run rate of 7 runs per over.

Team A is unable to complete their 20 overs quota. Team A sets a target score of 110 runs in 20 overs for team B.

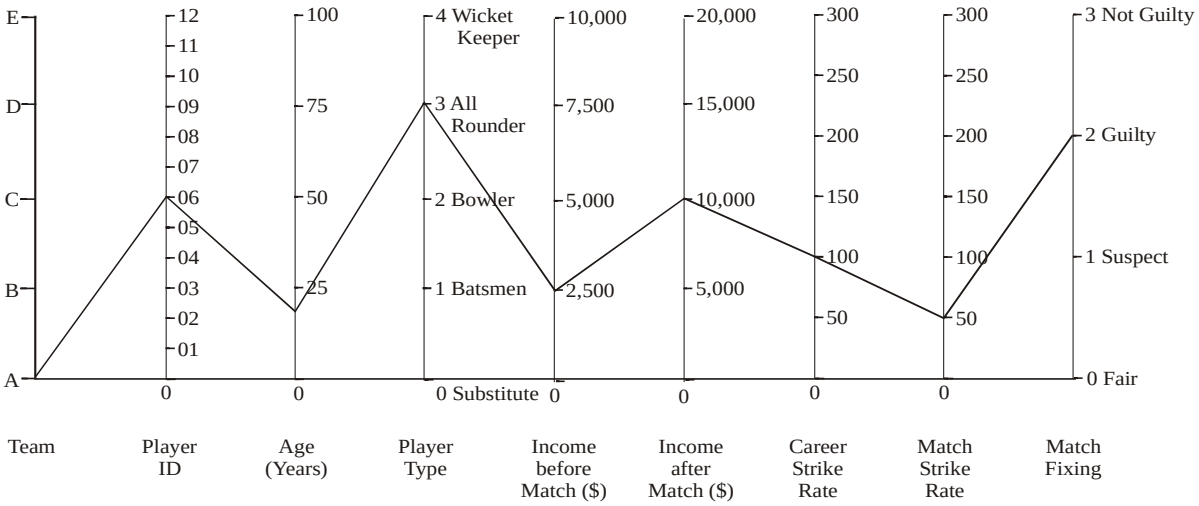


Graph 2: Iceberg Diagram for Team B (Target Chaser) in Match 1

Graph 2 represents Iceberg diagram for team B as target chaser. In its first five overs(overs 1-5) team B scores 75 runs by loosing 1 wicket at an average of 15 runs per over.

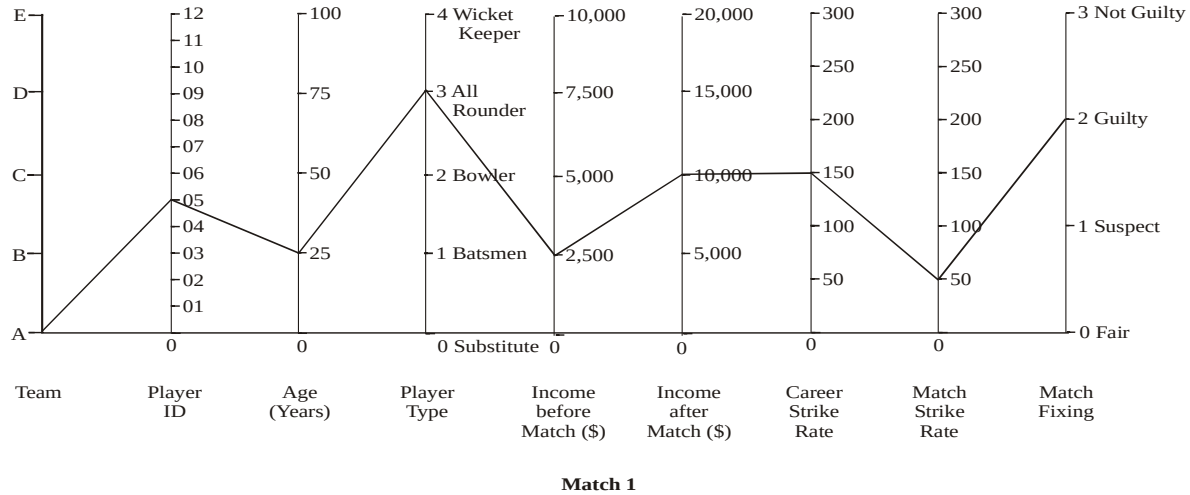
In its next five overs(overs 6-10) team B scores 35 runs at an average run rate of 7 runs per over by loosing 2 wickets. They are above the required target run rate. Team B in their first ten overs are able to score the target set by team A and wins the match.

Match 1 between Team A and Team B in Graph 4 and graph 5 represents Iceberg diagrams for players involved in match fixing. Here Iceberg diagram is used to represent favors received by players . These Iceberg diagrams include important dimensions like team, player id, age, player type(batsmen, bowler, all rounder, wicket keeper), income before match and income after match, player's career strike rate and match strike rate and match fixing status of player(fair, suspect, guilty, not guilty).



Graph 3: Iceberg Diagram for Match 1 Representing Player ID 06 involved in Match Fixing

In graph 3 Iceberg diagram shows how player with player id 06 belonging to team A, aged 20 years, who is an all rounder whose income before match is $ 2500 and income after match is $ 10000. His career strike rate is 100 but his current match run rate is 50 and he is found to be guilty of match fixing.



Graph 4: Iceberg Diagram for Match 1 Representing Player ID 05 involved in Match Fixing

Graph 1 represents Iceberg diagram for team A as target setter in match 1 between team A and team B. Team A in its first 5 overs (overs 1-5) of its batting innings scores 50 runs by loosing 4 wickets at an average run rate of 10 runs per over.

In the next five overs(overs 6-10), team A scores 25 runs by loosing 5 wicket at an average run rate of 5 runs per over.

In the next five overs (overs 11-15), team A scores 35 runs by loosing 1 wicket at an average run rate of 7 runs per over.

Team A is unable to complete their 20 overs quota. Team A sets a target score of 110 runs in 20 overs for team B.

In a T20 tournament involving 3 teams A, B and C and in match 1 between team A and team B, team A wins the toss and elects to bat. According to pitch report and weather report, the minimum expected target to be set is 225. Also in the tournament team A is the highest ranked team (Rank 1) followed by team C (Rank 2) and team B (Rank 3). In match 1 team A as a target setter sets team B a target of 110 runs. In scoring 110 runs it utilizes 15 overs. Team B achieves this target in 10 overs. According to CIS finding this match is fixed by team A for the following reasons

1. Even though team A is a higher ranked team it has under-performed.
2. It has not utilized all the 20 overs allotted to it.
3. It has lost wickets regularly and could not utilize last 5 overs.
4. The target score set is a very mere target when compared to expected minimum target.
5. Team A did not want to face Team C in the tournament final. For this reason it has lost to team B so that now it can face team B in the tournament final.

All these information can be inferred from Iceberg diagram shown in the result for match 1. To know the culprits involved in match fixing dimensions like income of player before match and income of player after match, players career strike rate and match strike rate are included.

Strike rate comparison helps in identifying players under performance. Income before match and income after match indicates illegal financial transactions done by the player for match fixing.

Both these dimensions clearly help to identify match fixing.

XII. CONCLUSION

Today Match fixing in cricket is a ground reality. Session fixing is a new form of match fixing where betting and match fixing in done in different sessions of the match and it provides bookies more opportunity for betting and detection of session fixing is much more complicated than match fixing. Multi relational clustering helps in analyzing multi dimensional data that deals with match fixing. Key challenges in application of multi relational clustering are highlighted in this paper along with Simpson's Paradox.

XIII. FUTURE ENHANCEMENTS

For multi dimensional data analysis new techniques like relevance analysis, dimensionality reduction techniques can be incorporated. Nowadays new types of match fixing like spot fixing and bracket fixing are used by bookies. So there is a need for developing solutions for these type of fixing problems. If the captain himself is involved in match fixing then it leads to bracket fixing. Bracket fixing is more complicated than session fixing.

XIV. ACKNOWLEDGEMENT

Authors wish to thank the Management of Adarsha Institute of Technology, Bangalore and special thanks to Dr.P.V.Krupakara for their constant support.

REFERENCES

[1] Anany Levitin, "Introduction to the Design and Analysis of Algorithms", Pearson Publications

[2] G.K.Gupta, "Introduction to Data Mining with case studies", PHI Learning Private Limited.

[3] R.V.Hauck et.al, "Using coplink to analyze criminal justice data", Computer, March, 2002, 30-37.

[4] Alex Berson, Stephan J Smith " Data Warehousing, Data Mining and OLAP", Tata McGraw Hill Edition publication.

[5] Pang-Ning Tan, Michael Steinbach and Vipin kumar "Introduction to Data Mining", Pearson Education publication

[6] Jiawei Han and Micheline Kambe "Data Mining Concepts and Techniques", 2ND Edition,Morgan Kaufmann publishers An imprint of Elsevier.

[7] K.Udayakumar and K.Ananthapadmanabha, KL Cluster Nearest Neighbor Outlier Prediction Algorithm for Match Fixing in Cricket, International Journal of Advances in Electronics and Computer Science, Special Issue, Sep.-2016, pp. 79-81, ISSN: 2393-2835.

[8] K.Udayakumar and K.Ananthapadmanabha, Algorithmic Design Notation for Match Fixing in Cricket Using Outlier Analysis, Higher Education Conclave 2016, p. 31, 2016, ISBN No. 978-81-8281-575-9.

[9] K.Udayakumar and K.Ananthapadmanabha, Forest Fire Model Proposal for Match Fixing in Cricket Based on Criminal Network Analysis, International Journal of Engineering Research, Volume No.5 Issue: Special 4, pp: 790-991, May 2016, doi: 10.17950/ijer/v5i4/016, ISSN: 2319-6890 (online), 2347-5013(print).

[10] G. Wang, H. Chen and H. Atabakhsh "Automatically detecting deceptive criminal identities ", Comm. ACM, March 2004, pp 70-76.

[11] T. Senator et al, " The FinCEN Artificial Intelligence system Identifying potential money laundering from reports of large cash transactions ", AI magazine vol. 16, no. 4, 1995, pp 23-39.

[12] K.Ananthapadmanabha and Dr.K.Udayakumar, Match Fixing Network Analysis to Verify Nearness among Internal Participants of a Cricket Match, 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, at SEVC, Bangalore, IEEE xplore ISBN: 978-1-5090-3704-9.

[13] K.Ananthapadmanabha and Dr.K.Udayakumar, Role of Iceberg Diagram as a Data Visualization Tool for Verifying Match Fixing in Cricket, 2017, 3rd National Conference on Recent Trends in Computer Science and Engineering (NCRTCS), at SJCIT, Bangalore on 3rd and 4th May, 2017, ISBN 978-81-931545-0-2.

[14] K.Ananthapadmanabha and Dr.K.Udayakumar, "Trustworthy Collaborative Investigation System for Match Fixing in Cricket Using Spectator Voting Scheme" at International Conference on Signal Processing Communication and Automation (ICSIPCA), 2017 in Tata McGraw Hill Journal, Grenze Digital Library and Grenze International Journal of Engineering and Technology (GIJET) ISSN: 2395-5295 (Online) 2395-5287 (Print) Indexed in Scopus (Cross Ref. or other indixing services).