A Data Mining Approach on Cluster Analysis of IPL

Pabitra Kumar Dey, Gangotri Chakraborty, Purnendu Ruj, and Suvobrata Sarkar

***Abstract*—Fuzzy clustering is an important approach in data mining. It has been applied broadly in many aspects and receiving great attention from enterprisers and scholars. This paper makes use of MATLAB language to produce a fuzzy clustering algorithm for classifying the batting statistics of Indian Premier League (IPL) T-20 version-3 cricket tournament into several numbers of clusters. The definition of clusters as well as the membership function has been implemented using MATLAB. The results obtained from Indian premier league batting statistics dataset detect n-clusters to handle the imprecise and ambiguous result. Finally, this article proposed a fuzzy clustering technique which provides efficient and accurate data analysis in the field of data mining.**

***Index Terms*—Cluster analysis, fuzzy set theory, machine learning, data mining.**

I. Introduction

Cluster analysis is a technique which discovers the substructure of a data set by dividing it into several clusters. Clustering plays an important role in data analysis and interpretation. It has been widely used for data analysis and has been an active subject in several research fields such as statistics, pattern recognition and machine learning. In the context of machine learning, clustering is an unsupervised learning method that groups' data into subgroups called clusters based on a well defined measure of similarity between two objects. Such kind of cluster-represented data provides a simpler description of the original data set but without loss of much information. A variety of clustering approaches have been developed for different goals and applications in specific areas. More comprehensive reviews of clustering approaches and clustering related issues can be found in [1]-[5].

Fuzzy clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships. Introducing fuzziness to clustering gives us the flexible representations of substructures of the data set. In 1965, L.A. Zadeh discovered fuzzy sets and systems in order to exploit the tolerance of imprecision, partial truth, and uncertainty to achieve robustness, tractability at low cost solution [6], [7]. There are different shapes of cluster centers and prototypes. Most of them conduct clustering in accordance with similarity or dissimilarity derived from distances, from the centroid of the cluster to data points. Lee et. al. [8] presents a new iterative fuzzy clustering algorithm that incorporates a supervisory scheme into an unsupervised fuzzy clustering process.

A fuzzy clustering approach for the classification of cosmetic defects is presented [9]. Hichem Frigui et. al. introduce a semi-supervised approach for clustering and aggregating relational data (SS-CARD) [10].

Data mining is the task of discovering interesting and hidden patterns from large amounts of data where the data can be stored in databases, data warehouses, on-line analytical process or other repository information [11]. It is also defined as knowledge discovery in databases (KDD) [12]. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, neural networks, information retrieval, etc [13]. Yaonan Wang et. al. proposes a center initialization approach based on a minimum spanning tree to keep FCM from local minima [14].

Indian Premier League (IPL) is a Twenty20 cricket competition initiated by the Board of Control for Cricket in India (BCCI) headquartered in Mumbai. It was started from 2008 consisting of 8 teams (franchises), where cricket players from different countries can participate. Since then IPL has become very popular throughout the world-wide. On 21 March 2010, at Chennai it was announced that for IPL 4th edition, two new teams from Pune and Kochi will be added. This will increase the number of franchises from 8 to 10 and the number of matches from 60 to 94 if the same format is used. In this paper, IPL3 bating statistics records have been considered for cluster analysis which is readily available from IPL website.

We proposed a fuzzy clustering technique to handle the imprecise and unambiguous data. N-clusters have been detected from IPL dataset. To define the membership function and threshold equation, MATLAB has been used and also to measure the distance between the centroid of the clusters and the several points. Finally, a decision is to be taken whether the corresponding point belongs to Cluster 1, Cluster 2 to N-clusters or neither belongs into any cluster.

The paper is organized as follows: Section 2 discuss about the various issues of Cluster Analysis. Section 3 focuses about the basic concepts of data mining. Section 4 represents the design of fuzzy database taking IPL dataset into account. Experiment and results are carried out on section 5. Finally, section 6 concludes the paper.

Manuscript received May 21, 2012, revised June 21, 2012.

Pabitra Kumar Dey is with Department of Computer Application, Dr.B.C.Roy Engineering College, Durgapur, India (Email-dey_pabitra@yahoo.co.in)

Gangotri Chakraborty is with Manipal University, India.

Purnendu Ruj is with Department of Computer Science and Engineering, Dr.B.C.Roy Engineering College, Durgapur, India.

Suvobrata Sarkar is with Department of Computer Science and Engineering, Dr.B.C.Roy Engineering College, Durgapur, India.

II. Cluster Analysis

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A loose definition of clustering could be "the process of

organizing objects into groups whose members are similar in some way".

A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [15].

We can show this with a simple graphical example:

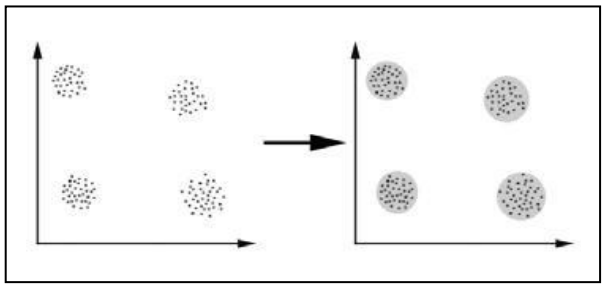

Fig. 1. Example of cluster

In this case, we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

A. Cluster Parameters

In this paper, clusters have been described by the following parameters:

Centroid – Defined the centre of gravity for all clusters members

$$C_m = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

Distance between two Centroid – distance between two clusters centers

$$d_{mn} = \|C_m - C_n\| \tag{2}$$

Manhattan Distance Measure –

$$d_1(x_i, x_j) = \left|x_{i,k} - x_{j,k}\right| \tag{3}$$

B. The Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties (natural" data types), for useful and suitable groupings ("useful" data classes) or unusual data objects (outlier detection).

C. Applications

Clustering algorithms can be applied in many fields, for instance:

- *Marketing*: finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records;
- *Biology*: Classification of plants and animals given their features;
- *Libraries*: Book ordering;
- *Insurance*: Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- *City-planning*: Identifying groups of houses according to their house type, value and geographical location;
- *Earthquake studies*: Clustering observed earthquake epicentres to identify dangerous zones;
- *WWW*: Document classification; clustering weblog data to discover groups of similar access patterns.

III. Data Mining

Data mining is basically a concept and can be considered as a part of knowledge discovery in databases (KDD). This process consists mainly of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation. Association rule techniques are used for data mining if the goal is to detect relationships or associations between specific values of categorical variables in large data sets. Data mining is the process of discovering meaningful patterns and relationships that lies hidden within very large databases [16]. Apart from these, data mining as the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [17].

The architecture of a typical data mining system may have the following major components: database, data warehouse, or other information repository; a server which is responsible for fetching the relevant data based on the user's data mining request, knowledge base which is used to guide the search[13]. Data mining engine consists of a set of functional modules, Pattern evaluation module which interacts with the data mining modules so as to focus the search towards interesting patterns and graphical user interface which communicates between users and the data mining system, allowing the user interaction with system.

IV. IPL Dataset

The concept of clustering has been considered in order to classify the IPL3 bating statistics of fuzzy data into appropriate clusters. A fuzzy database has been constructed by using MATLAB and the insertion of whole records comprises of 181 data which are collected from IPL website. The dataset consists of several attributes like player-id, player name, team name, innings, runs, average, balls and strike rates which are clearly shown in Fig. 2.

ID	Player	Team	Inns	Runs	Avg	Balls	SR
144	A Kumble	RCB	5	6	-	11	54.55
58	A Mishra	DD	7	39	9.75	51	76.47
145	A Mithun	RCB	1	5	5	4	125
61	A Nehra	DD	2	23	23	20	115
42	A Symonds	DC	16	429	30.64	341	125.81
177	A Uniyal	RR	2	4	4	7	57.14
25	AA Bilakhia	DC	1	2	2	4	50
162	AA Jhunjhunwala	RR	11	183	20.33	166	110.24
67	AB Agarkar	KKR	4	40	40	29	137.93
93	AB Barath	KXIP	3	42	21	42	100
47	AB de Villiers	DD	7	111	15.86	120	93.28
69	AB Dinda	KKR	1	0	0	1	0
57	AB McDonald	DD	3	65	-	52	125
27	AC Gilchrist	DC	16	289	18.06	185	156.22
178	AC Voges	RR	7	181	45.25	143	126.57
166	AD Mascarenhas	RR	2	12	12	10	120
80	AD Mathews	KKR	11	233	33.29	184	126.63
126	AG Murtaza	MI	0	0	-	0	-
172	AG Paunikar	RR	1	0	0	1	0
161	AJ Finch	RR	1	21	21	21	100
127	AM Nayar	MI	3	58	29	51	113.73
114	AN Ahmed	MI	1	4	-	7	57.14
24	Anirudh Singh	DC	4	63	15.75	66	95.45
159	AP Dole	RR	2	34	17	22	154.55
131	AP Tare	MI	4	51	12.5	36	138.89

Fig. 2. Snapshots of IPL3 batting dataset

V. Experiment and Result

The membership function is generated from the fuzzy database corresponding to each record into values which lies in the range of 0 to 1 taking runs/balls as parameter. The membership function and its corresponding graph are shown in Fig.3 and the graph of player id and the membership value are shown in the Fig.4.

To define the fuzzy membership function, MATLAB has been used whose value varies from 0 to 1.

Membership Function:-

$$\mu(x)=\begin{cases}0 & ,x\le a\\ \dfrac{x-a}{b-a} & ,a\le x\le b\\ \dfrac{c-x}{c-b} & ,b\le x\le c\\ 0 & ,x\ge c\end{cases}$$

where x = runs/balls, a = minimum of x, b = median of x and c = maximum of x. (4)

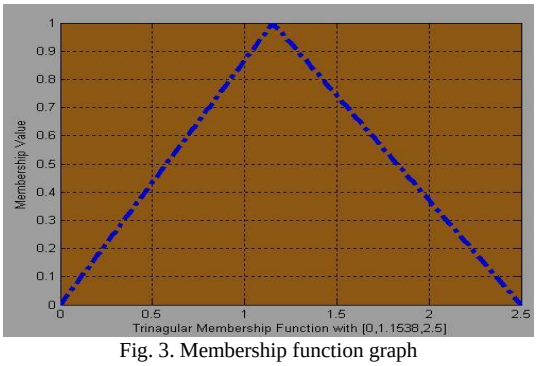

Fig. 3. Membership function graph

***Centroid of k^th cluster*:-**

Centroid$_k$ = k/(n+1) where k = 1, 2, … , n , where n = Number of cluster (5)

Threshold Equation:- Threshold = standard deviation of x/(n-1) (6)

Range of Cluster k:- Range = (Centroidk – Threshold, Centroidk + Thresholod) (7)

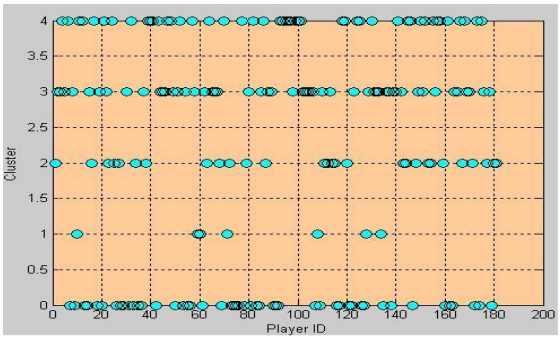

Fig. 4. Player id vs. Membership value graph

Now we consider for n = 4 then cluster-1 to cluster-4 are present and we try to segregate the data into this four cluster and the data which are not belong to that cluster and the value of that cluster = 0. The graph of player id and the corresponding cluster for the cluster =4 are shown in the Fig. 5. Out of 181 records the proposed technique classifying into several cluster is 133 i.e. of 73.48% accuracy in the case of cluster = 4. The snapshots of cluster detection and the corresponding membership values of IPL players are displayed in Fig.7 and the corresponding MATLAB implementation are shown in the Fig.6.

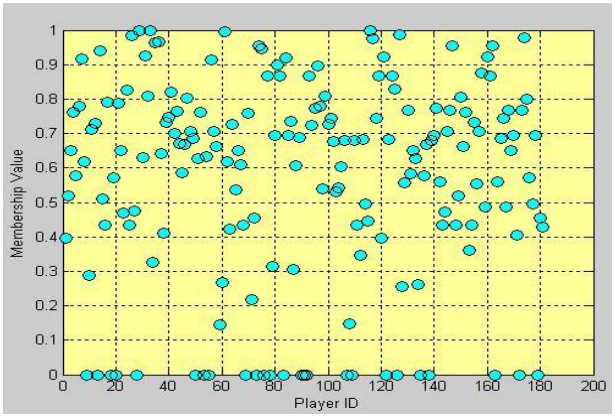

Fig. 5. Player ID vs. Cluster graph

```
k = (6+n+1);
for i=1:n
    for j=1:181
        if((ipl3(j:j,i+6)<threshold))
            ipl3(j:j,k)=i;
        end
    end
end
xlswrite('C:\Program
Files\MATLAB71\work\ipl3_batting.xls',ipl3,'Cluster');
for i=1:n
    c(i)=0;
    for j=1:181
        if(ipl3(j:j,k)==i)
            c(i)=c(i)+1;
        end
    end
end
disp('The Total no. of Records belong in each
Cluster.....................');
disp(c);
plot(ipl3(:,1),ipl3(:,k),'ko','MarkerEdgeColor','k',...
'MarkerFaceColor','c','MarkerSize',8);
hgsave('batting_cluster3');
```

Fig .6. MATLAB implementation for clustering

ID	Player	Team	Inns	Runs	Avg	Balls	SR	Distance-1	Distance-2	Distance-3	Distance-4	Cluster
144	A Kumble	RCB	5	6	-	11	54.55	0.2727273	0.072727273	0.12727273	0.32727273	Cluster-2
58	A Mishra	DD	7	39	9.75	51	76.47	0.4627451	0.262745098	0.0627451	0.1372549	Cluster-3
145	A Mithun	RCB	1	5	5	4	125	0.5057143	0.305714286	0.10571429	0.09428571	Cluster-4
61	A Nehra	DD	2	23	23	20	115	0.7966667	0.596666667	0.39666667	0.19666667	
42	A Symonds	DC	16	429	30.64	341	125.81	0.4997235	0.299723502	0.0997235	0.1002765	
177	A Uniyal	RR	2	4	4	7	57.14	0.2952381	0.095238095	0.1047619	0.3047619	Cluster-2
25	AA Bilakhia	DC	1	2	2	4	50	0.2333333	0.033333333	0.16666667	0.36666667	Cluster-2
162	AA Jhunjhunwala	RR	11	183	20.33	166	110.24	0.7554217	0.555421687	0.35542169	0.15542169	
67	AB Agarkar	KKR	4	40	40	29	137.93	0.4096552	0.209655172	0.00965517	0.19034483	Cluster-3
93	AB Barath	KXIP	3	42	21	42	100	0.6666667	0.466666667	0.26666667	0.06666667	Cluster-4
47	AB de Villiers	DD	7	111	15.86	120	93.28	0.6016667	0.401666667	0.20166667	0.00166667	Cluster-4
69	AB Dinda	KKR	1	0	0	1	0	0.2	0.4	0.6	0.8	
57	AB McDonald	DD	3	65	-	52	125	0.5057143	0.305714286	0.10571429	0.09428571	Cluster-4
27	AC Gilchrist	DC	16	289	18.06	185	156.22	0.2738224	0.073822394	0.12617761	0.32617761	Cluster-2
178	AC Voges	RR	7	181	45.25	143	126.57	0.494026	0.294025974	0.09402597	0.10597403	Cluster-3
166	AD Mascarenhas	RR	2	12	12	10	120	0.5428571	0.342857143	0.14285714	0.05714286	Cluster-4
80	AD Mathews	KKR	11	233	33.29	184	126.63	0.4936025	0.293602484	0.09360248	0.10639752	Cluster-3
126	AG Murtaza	MI	0	0	-	0	-	0.2	0.4	0.6	0.8	
172	AG Paunikar	RR	1	0	0	1	0	0.2	0.4	0.6	0.8	
161	AJ Finch	RR	1	21	21	21	100	0.6666667	0.466666667	0.26666667	0.06666667	Cluster-4
127	AM Nayar	MI	3	58	29	51	113.73	0.7856209	0.585620915	0.38562092	0.18562092	
114	AN Ahmed	MI	1	4	-	7	57.14	0.2952381	0.095238095	0.1047619	0.3047619	Cluster-2
24	Anirudh Singh	DC	4	63	15.75	66	95.45	0.6272727	0.427272727	0.22727273	0.02727273	Cluster-4
159	AP Dole	RR	2	34	17	22	154.55	0.2862338	0.086233766	0.11376623	0.31376623	Cluster-2
131	AP Tare	MI	4	51	12.5	36	138.89	0.3819048	0.181904762	0.01809524	0.21809524	Cluster-3

Fig. 7. Membership values and cluster detection

VI. Conclusion

Data Clustering plays a major role in grouping the similar type of data into a specific cluster. Cluster analysis aims at identifying groups of similar objects and, therefore helps to discover distribution of patterns and interesting correlations in large data sets. Fuzzy clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships. In this paper, fuzzy clustering has been adopted using fuzzy relational database to detect two clusters on the IPL3 batting statistics dataset. The records in the database are partitioned in a manner such that similar records are in the same cluster. N-clusters have been detected from IPL batting statistics dataset. MATLAB has been used for the definition the membership function, threshold equation and detecting the several clusters.

The future research work of this article lies on the fact that all version of IPL dataset will be taken into consideration to develop an algorithm which detects n-clusters to generalize the fuzzy clustering techniques and comparison of all version of IPL dataset.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review", *ACM Computing Surveys* 31 (3), pp: 264–323, 1999.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, second ed., Morgan Kaufmann, California, 2005.

[3] P. Berkhin, *Survey of clustering data mining techniques*, Technical Report, Accrue Software, Inc., 2002.

[4] R. Xu and D. Wunsch, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks* 16 (3) 645–678, 2005.

[5] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering", *Pattern Recognition* 41, 176–190, 2008.

[6] L. A. Zadeh, *Fuzzy sets: Information and Control*, 1965.

[7] L. A. Zadeh, *From search engines to question-answering systems the role of fuzzy logic*, University Berkeley, California.

[8] H. Lee, K. H. Park, and Zeungnam Zenn Bien, "Iterative Fuzzy Clustering Algorithm With Supervision to Construct Probabilistic Fuzzy Rule Base From Numerical Data",*IEEE Transactions on Fuzzy Systems*,Vol.16, No.1, June, 2008.

[9] M. M. Chacón and S. J.Nevarez, *A Fuzzy Clustering Approach on the Classification of Non Uniform Cosmetic Defects* .

[10] Hichem Frigui and Cheul Hwang, "Fuzzy Clustering and Aggregation of Relational Data With Instance-Level Constraints", *IEEE Transactionson Fuzzy Systems*, Vol.16, No.1, December, 2008.

[11] Maria Halkidi, "*Quality assessment and Uncertainty Handling in Data Mining Process*", [Online]. Available: http://www.edbt2000.uni-konstanz.de/phd-workshop/papers/Halkidi.pdf

[12] Fayyad, U. M., G. P. Shapiro, and P. Smyth. "From Data Mining to Knowledge Discovery in Databases", 0738-4602-1996, *AI Magazine* pp: 37–53, (Fall 1996).

[13] J. Han and M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Champaign: CS497JH, fall 2001, www.cs.sfu.ca/~han/DM_Book.html.

[14] Yaonan Wang, Chunsheng Li, and Yi Zuo, "A Selection Model for Optimal Fuzzy Clustering Algorithm and Number of Clusters Based on Competitive Comprehensive Fuzzy Evaluation", *IEEE Transactions on Fuzzy Systems*, Vol.17,No.3, June, 2009.

[15] Pabitra Kumar Dey, Gangotri Chakraborty, and Suvobrata Sarkar, "Cluster Detection Analysis using Fuzzy Relational Database", *ICCEE* 2010, Vol. 6, pp:84-87, China, 2010.

[16] Claude Seidman. "Data Mining with Microsoft SQL Server 2000 Technical, Reference", ISBN: 0-7356-1271-4, [online]. Available: www.amazon.com/Mining-Microsoft-Server-Technical-Reference/dp/0735612714.

[17] David Hand, Heikki Mannila, and Padhraic Smyth. "Principles of Data Mining", ISBN: 026208290, *MIT Press*,Cambridge, MA,2001.

Pabitra Kumar Dey is working as an Asst. Prof. in the Dept. of Computer Application, Dr. B.C.Roy Engineering College, Durgapur, India. He was born on 10/12/1978. He obtained B.Sc.(Math Hons.) in 2000, M.C.A. in 2004 & M.Tech.(CST) in 2011. He has about more than of 7 years of Teaching Experience and 3 years of Research Experience. He has more than 10 research papers in reputed journals and conference proceedings.

Gangotri Chakraborty obtained her M.tech(CST) degree from W.B.U.T. in 2011 and she is presently working in Sikkim Manipal University as a lecturer.

Purnendu Ruj obtained his M.tech(CST) degree from W.B.U.T. in 2011 and he is presently working in Dr. B.C.Roy Engineering College, Durgapur, India.

Suvobrata Sarkar is working as an Asst. Prof. in the Dept. of Computer Science & Engineering, Dr. B.C.Roy Engineering College, Durgapur, India. He has several research papers in reputed journals and conference proceedings.