Cluster Detection Analysis Using Fuzzy Relational Database

Pabitra Kumar Dey, Gangotri Chakraborty, and Suvobrata Sarkar

Abstract—**In order to handle imprecise and ambiguous information, the application of fuzzy set theory for the design of database, information storage and retrieval systems has been gaining popularity recently. This paper gives emphasis on the basic characteristics of fuzzy relational databases, their properties, along with the data clustering in database systems. Indian premier league dataset has been considered for the detection of clusters. Several clustering parameters like centroid, radius and Manhattan distance measure have been applied. The definition of clusters as well as the membership function has been implemented using PL/SQL. The results obtained from Indian premier league batting statistics dataset detect two clusters, namely Cluster 1 and Cluster 2. Finally, this article proposed a fuzzy database organization and clustering of records which provides efficient and accurate fuzzy retrieval.**

Index Terms—**Cluster analysis, fuzzy set theory, data mining, fuzzy relational database, information retrieval.**

I. Introduction

Cluster analysis is a technique which discovers the substructure of a data set by dividing it into several clusters. It is largely involved in data mining approaches. 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 [1], [2]. Fuzzy clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships. 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. Moreover, the important properties of fuzzy relational databases are preserved in a generalized model built on equivalence relations on finite database domains. Further, the notion of a functional dependency to the fuzzy relational model has been generalized. For representing fuzzy data, we propose a possibility distribution-fuzzy-relational model, in which fuzzy data are represented by fuzzy relations whose grades of membership and attribute values are of possibility distribution. The fuzziness of an attribute value is represented by possibility distribution and its association by a grade of membership. This model is an extension of the relational model of data [3], [4].

Manuscript received August 25, 2012; revised October 27, 2012.

Pabitra Kumar Dey is with the Department of Computer Application, Dr. B.C Roy Engineering College, Jemua Road, Fuljhore, Durgapur-713206 , West Bengal, India (e-mail: dey_pabitra@yahoo.co.in)

Gangotri Chakraborty is with the Department of Computer Science, Sikkim Manipal University, India (e-mail: gangotri1986@gmail.com)

Suvobrata Sarkar is with Department of Computer Science and Engineering, Dr. B.C Roy Engineering College, Jemua Road, Fuljhore, Durgapur-713206, West Bengal, India (e-mail: suvobrata.sarkar@ieee.org)

An alternative to the conventional fuzzy clustering algorithms which forms fuzzy cluster in order to minimize the total distance from cluster centers to data points [5]. Several fuzzy database prototypes have been developed since then by incorporating the rules of fuzzy set theory into conventional database models, making them capable of handling uncertain and imprecise data [6], [7]. A.K. Sharma et. al [8] proposed an algorithm for the discovery of fuzzy inclusion dependencies that may exists between two given fuzzy relations stored in one or more fuzzy relational database. Pickert et. al [9] developed a strategy for systematic verification and improvement for their contextual analysis by a fuzzy clustering approach using non-redundant libraries of search profiles as a prerequisite. Z.M.Ma et. al [10] discussed about the issues of fuzzy functional dependencies by proposing a set of sound inference rules, which are similar to Armstrong's axioms for classical cases of fuzzy functional dependencies. Shyue-Liang Wang et. al [11] extended the concept of fuzzy functional dependency to approximate dependency on similarity based fuzzy automatically obtained approximate answers for null queries and missing data values in an incomplete database. This kind of facility improved the cooperative nature of databases and enhanced the user friendliness of the database system. There are a number of very good surveys of fuzzy relational database approach [12], [13], [14]; we will not go into details here but just point to some of its essential contributions.

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 whole world. 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 by using the relational database. Two clusters have been detected from IPL dataset. To define the membership function and threshold equation, PL/SQL 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 or Cluster 2 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 fuzzy relational database. Section 4 represents the design of fuzzy database using Pl/SQL taking

DOI: 10.7763/IJIEE.2013.V3.306

IPL dataset into account. Experiment and results are carried out on section 5. Finally, section 6 concludes the paper.

II. Cluster Analysis

Cluster analysis aims at identifying groups of similar objects and, therefore helps to discover distribution of patterns and interesting correlations in large data sets [15], [16]. Data Clustering is a technique in which, the information that is logically similar is physically stored together. In clustering the objects of similar properties are placed in one class and a single access to the disk makes the entire class available. Clusters should be exhaustive and mutually exclusive so that all data points are in cluster forms and the cluster boundaries are well defined. It is also desirable that there should be a number of clusters in order to insure that all the cluster entries in the cluster table are not greater than the number of records in the database itself. Complimentary to this property, each cluster should contain as many applicable records as possible. This increases database access efficiency.

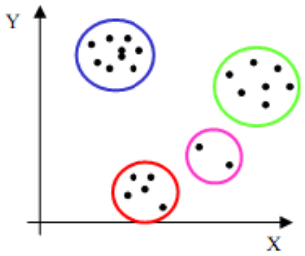

Fig. 1. Exemplary clusters

A. Cluster Parameters

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

Centroid – Defined the centre of gravity for all clusters members

$$c_m = \frac{\sum_{m}^{N} x_1}{N} \quad (1)$$

Radius – Representing average distance between all cluster members and cluster centre

$$R_m = \sqrt{\frac{\sum_{m}^{N} (x_1 - c_m)}{N}} \quad (2)$$

Distance between two Centroid – distance between two clusters centers

$$d_{mn} = \left\| C_M - C_n \right\| \quad (3)$$

Manhattan Distance Measure –

$$d_1(x_i - x_j) = \left| x_{i,k} - x_{j,k} \right| \quad (4)$$

B. Benefits of Clustering

Data Clustering plays the major role in every field of life, especially in data mining approaches. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. The use of data clustering has its own values, especially in the field of information retrieval. Fast information retrieval from the databases has always been a significant issue. There are a lot of major applications of clustering technique viz. to detect disease like tumor, the scanned pictures or the X-rays are compared with the existing ones and the dissimilarities are recognized.

III. Fuzzy Relational Database

Relational database is considered as a very successful model for a wide range of applications. A fuzzy relational database differs from the original relational model in two respects [17]. Firstly, attribute values need not contain any atomic value in a fuzzy database, and the values are restricted to be drawn from a single set (the corresponding domain of the attribute). Secondly, the identity relationship among attribute values in the original relational model is replaced with a similarity relationship that reflects the closeness of attribute values. A fuzzy database system must be capable of processing information stored in conventional database systems. A relational database consists of one or more relations [18]. Each relation consists of one or more attributes. In a fuzzy relational database, a relation R is characterized by following membership function:

$$UR : D_1 \times D_2 \times \ldots\ldots\ldots \times D_n \rightarrow \{0,1\} \quad (5)$$

where D_i (i = 1,2,. . . , n) is the domain for each attribute A_i of a relation R, and × denotes the Cartesian product.

IV. IPL Dataset

ID	Player	Team	Inns	Runs	Avg	Balls	SR
1	A Kumble	RCB	5	6	-	11	54.55
2	A Mishra	DD	7	39	9.75	51	76.47
3	A Mithun	RCB	1	5	5	4	125
4	A Nehra	DD	2	23	23	20	115
5	A Symonds	DC	16	429	30.64	341	125.81
6	A Uniyal	RR	2	4	4	7	57.14
7	AA Bilakhia	DC	1	2	2	4	50
8	AA Jhunjhunwala	RR	11	183	20.33	166	110.24
9	AB Agarkar	KKR	4	40	40	29	137.93
10	AB Barath	KXIP	3	42	21	42	100
11	AB de Villiers	DD	7	111	15.86	119	93.28
12	AB Dinda	KKR	1	0	0	1	0
13	AB McDonald	DD	3	65	-	52	125
14	AC Gilchrist	DC	16	289	18.06	185	156.22
15	AC Voges	RR	7	181	45.25	143	126.57
16	AD Mascarenhas	RR	2	12	12	10	120
17	AD Mathews	KKR	11	233	33.29	184	126.63
18	AG Murtaza	MI	0	0	-	0	-
19	AG Paunikar	RR	1	0	0	1	0
20	AJ Finch	RR	1	21	21	21	100
21	AM Nayar	MI	3	58	29	51	113.73
22	AN Ahmed	MI	1	4	-	7	57.14
23	Aniruth Singh	DC	4	63	15.75	66	95.45
24	AP Dole	RR	2	34	17	22	154.55
25	AP Tare	MI	4	50	12.5	36	138.89

Fig. 2. Snapshots of IPL3 batting 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 PL/SQL 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.

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

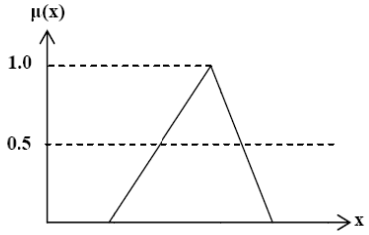

Fig. 3. Membership function

ID	Player	Team	Inns	Runs	Avg	Balls	SR	Grade	Cluster1
1	A Kumble	RCB	5	6	-	11	54.55	0.2727273	
2	A Mishra	DD	7	39	9.75	51	76.47	0.3823529	c-1
3	A Mithun	RCB	1	5	5	4	125	0.625	c-2
4	A Nehra	DD	2	23	23	20	115	0.575	
5	A Symonds	DC	16	429	30.64	341	125.81	0.6290323	c-2
6	A Uniyal	RR	2	4	4	7	57.14	0.2857143	
7	AA Bilakhia	DC	1	2	2	4	50	0.25	
8	AA Jhunjhunwala	RR	11	183	20.33	166	110.24	0.5512048	
9	AB Agarkar	KKR	4	40	40	29	137.93	0.6896552	c-2
10	AB Barath	KXIP	3	42	21	42	100	0.5	c-1
11	AB de Villiers	DD	7	111	15.86	119	93.28	0.4663866	c-1
12	AB Dinda	KKR	1	0	0	1	0	0	
13	AB McDonald	DD	3	65	-	52	125	0.625	c-2
14	AC Gilchrist	DC	16	289	18.06	185	156.22	0.7810811	
15	AC Voges	RR	7	181	45.25	143	126.57	0.6328671	c-2
16	AD Mascarenhas	RR	2	12	12	10	120	0.6	
17	AD Mathews	KKR	11	233	33.29	184	126.63	0.6331522	c-2
18	AG Murtaza	MI	0	0	-	0	-	0	
19	AG Paunikar	RR	1	0	0	1	0	0	
20	AJ Finch	RR	1	21	21	21	100	0.5	c-1
21	AM Nayar	MI	3	58	29	51	113.73	0.5686275	
22	AN Ahmed	MI	1	4	-	7	57.14	0.2857143	
23	Aniruddh Singh	DC	4	63	15.75	66	95.45	0.4772727	c-1
24	AP Dole	RR	2	34	17	22	154.55	0.7727273	
25	AP Tare	MI	4	50	12.5	36	138.89	0.6944444	c-2

Fig. 4. Membership values and cluster detection.

V. Experiment and Results

After the insertion of total records consisting of 181 IPL data using PL/SQL, 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. We have classified the membership values that are between (0.38 and 0.48) as cluster1 and those between (0.64 and 0.74) as cluster2. The membership values lying within the range 0.48 to 0.64 have an uncertainty association between them and might belong to either of the two clusters. The membership function and its corresponding graph are shown in Fig.3. The snapshots of cluster detection and the corresponding membership values of IPL players are displayed in Fig.4.

Membership Function:-

$$\mu(x) = \begin{cases} 0, & 0 \le x < 0.32 \\ x/2, & 0.32 \le x \le 1.85 \\ 1, & x > 1.85 \end{cases} \quad (6)$$

where x = Runs/Balls.

Threshold equation:-

Threshold = (((Centroid 2 – Centroid 1)/2)/2)

or,

Threshold= ((Centroid2 –Radius2)-(Centroid1+ Radius1))/2) (7)

where Radius1 and Radius2 are radius of Cluster 1 and Cluster 2 respectively.

PL/SQL has been implemented to apply K-Mean Clustering and Manhattan Distance formula to determine the distance between the fuzzy points and centroid of each of the two clusters and if the distance falls within the threshold value then a decision is taken for cluster detection. Accordingly, the uncertain intermediate membership values are classified into appropriate clusters. The corresponding graph for cluster 1 and cluster 2 are available in fig.5.

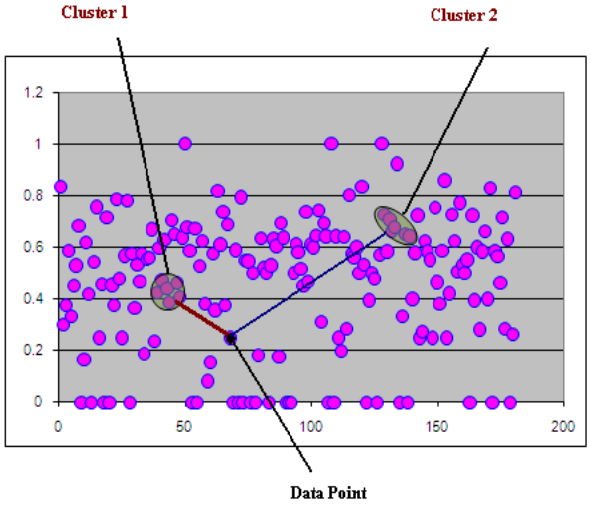

Fig. 5. Graph with two clustering

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. Two clusters, namely Cluster 1 and Cluster 2 have been detected from IPL batting statistics dataset. PL/SQL has been used for the definition the membership function and threshold equation.

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.

REFERENCES

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

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

[3] U. T. Imada, I. Hatono, and H. Tamura, "Fuzzy Retrieval from Fuzzy Object Oriented Database," in *Proc. of 10th Fuzzy System Symposium*, June 1994, Osaka Japan, pp. 93-96.

[4] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the AGM*, vol. 13, 1970, pp. 377-387.

[5] A. T. Grissa, "An Alternative Extension of the FCM Algorithm for Clustering Fuzzy Databases," in *Proc. of 2010 2nd International Conference on Advances in Databases, Knowledge, and Data Applications*, 2010.

[6] B. P. Buckles and F. E. Petry, "A Fuzzy Representation of Data for Relational Databases," *Fuzzy Sets and Systems*, vol. 7, 1982, pp. 213-226.

[7] B. P. Buckles and F. E. Petry, "Uncertainty Models in Information and Database Systems," *Journal of Information Science*, vol. 11, 1985, pp. 77-87.

[8] A. K. Sharma, A. Goswami, and D. K. Gupta, "Discovery of Fuzzy Inclusion Dependencies in Fuzzy Relational Databases," 2004.

[9] L. Pickert, I. Reuter, F. Klowonn, and E. Wingender, "Transcription regulatory region analysis using signal detection and fuzzy clustering," *Bio-informatics*, vol. 14 no. 3, 1998, pp-244-251.

[10] Z. M. Ma, W. J. Zhang, W. J. Ma, and G. Q. Chen, "Functional Dependencies in Extended Possibility-Based Fuzzy Relational Databases," 2000.

[11] S. L. Wang, T. P. Hong, "Mining Approximate Dependency to Answer Null Queries on Similarity-based Fuzzy Relational Databases," 2000.

[12] M. Umano, T. Imada, I. Hatono, and H. Tamura, "Implementation of a Fuzzy Object-Oriented Database," in *Proc. of* Sixth *International Fuzzy Systems Association World Congress*, pp. 401-404.

[13] S. M. Chen and W. T. Jong, "Fuzzy Query Translation for Relational Database Systems," *IEEE Transactions on Systems, Man, AND Cybernetics*—Part B, vol. 27, no. 4, August, 1997.

[14] H. K. Tripathy, B. K. Tripathy, P. K. Das, and S. P. Khadanga, "Application of Parallelism SQL in Fuzzy Relational Databases," *International Conference on Computer Science and Information Technology*, 2008.

[15] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure. I linear structure, fuzzy c-lines," *SIAM Journal of Applied Mathematics*, vol. 40, no. 2, 1981, pp 339-357.

[16] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure. II fuzzy c-varieties and convex combinations thereof," *SIAM Journal of Applied Mathematics*, vol. 40, no. 2, 1981, pp. 358-372.

[17] Q. Yang, W. Zhang, J. Wu, and H. Nakajima, "Efficient Processing of Nested Fuzzy SQL Queries in a Fuzzy Database," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, December, 2001.

[18] J. Kacprzyk, "Computing in words in database quering and data mining," *System Research Institute Polish Academy of Science*, Poland.

Mr. 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.

Miss 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.

Mr. 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.