Performance Prediction of Players in Sports League Matches

Praveen Kumar Singh[1], Muntaha Ahmad[2]

[1]M.Tech Student BIT Mesra Ranchi- 835215, Jharkhand, India

[2]Assistant Professor BIT Noida 201301 U.P., India

Abstract: ***The objective of this article is to discover the better performing team in the Hockey India League (HIL) for the purpose of formation of the winning team based on cluster analysis of their past performance by using the machine learning techniques. Two most prevalent machine learning techniques k-means and fuzzy clustering have been used respectively to predict the better performing player. The results of the two techniques proposed were compared and were found nearly identical. The complexity of initializing K-means clustering technique is resolved by using MacQueen algorithm. The results obtained from Hockey Indian League Goal statistics dataset were used to detect n-clusters to handle the imprecise and ambiguous result. Finally, this article proposed a K-Means clustering technique which provides efficient and accurate data analysis in the field of data mining.***

Keywords: Sports data mining, K-means clustering, fuzzy clustering, MacQueen algorithm, HIL.

1. Introduction

Cluster analysis is a method which explores the substructure of a data set by dividing it into many clusters. 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. Numerous clustering approaches have been developed for different goals and applications in specific areas [1, 2, 4, 6, and 8].

K-means is an unsupervised learning algorithm that provides solution for the well-known clustering problem. It uses simple means of classifying a given data set through a certain number of clusters fixed in advance. It defines k centroids, one for each cluster. These centroids are placed in an efficient way because of different location causes different result. It is a better method to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2$$

$$\left\|x_i^{(j)} - c_j\right\|^2$$

Where is a chosen distance measure between a data point and the cluster center is an indicator of the distance of the *n* data points from their respective cluster centers [7].

We used K means technique to handle the imprecise and unambiguous data. N-clusters have been detected from HIL dataset. 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 5 represents the design of HIL dataset. Experiment and results are carried out on section 6. Finally, section 7 concludes the paper.

Hockey India League (HIL) is a hockey competition initiated by HOCKEY INDIA [10]. It was started from 2013 consisting of 5 franchises, where hockey players from different countries can participate. Since then HIL has become very popular throughout the world-wide., a new team KALINGA LANCERS joined further. This will increase the number of franchises from 5 to 6. In this paper, HIL 2014 statistics records have been considered for cluster analysis which is readily available from HIL website.

1.1 Cluster Analysis

The definition of clustering may be considered as "the process of organizing objects into groups whose members are similar in some ways". A cluster is therefore a collection of objects which are "similar" between them and "dissimilar" to the object belonging to other clusters [3]. The following graph exhibits the illustrated concept.

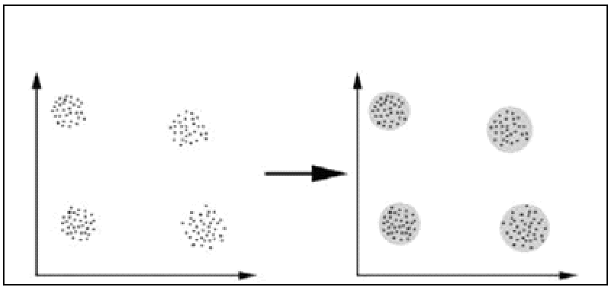


Fig. 1. Example of cluster

In this case, we easily identify 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.

1.2 Cluster Parameters

Clusters are characterized using the parameters described below.

Centroid- The center of gravity for all clusters members are defined as below [9].

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

Distance between two Centroid- distances between two clusters centers [7]

2. Predictive Analysis in Sports

Sports analytics is perceived as "the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision makers and enable them to help their organizations in gaining a competitive advantage on the field of play." For this purpose one has look into the increasingly diverse and sophisticated sources of data that in turn are driving explosive growth in the field of sports analytics. There is a need to examine the ways in which predictive models and information delivery systems are leveraging these growing mountains of data to create the types of competitive advantage that every team is after.

Predictive models are a key component of every effective sports analytics program because these models translate raw data into useful information. For example, there is very little value in the motion capture data described in Part 1 without skilled analysis that transforms that data from millions of raw records into actionable information a decision-maker can understand and trust. Data sources alone are clearly unusable, as no decision-maker would be able to draw conclusions from these mountains of raw data. Once such data is analyzed, however, the analysis results have the potential to become a valuable and unique tool to aid decision-makers in making better decisions.

Teams often begin their use of predictive analytics because they are looking for a tool to reduce the seemingly high error rate in decision-making around their sport's amateur draft. Given the financial investment and opportunity cost associated with high-round draft choices, teams that spend a draft pick on a player who does not make a significant contribution find they have made an expensive mistake. Conversely, drafting well can make a huge impact on a team's fortunes, sometimes immediately and often for several seasons to come.

3. Sports League Matches

3.1 Introduction

A sports league match is a group of sporting teams or individual players or athletes that compete against each other in a specific sport. At its simplest, it may include international players as well as be a local group of amateur athletes who form teams among themselves and compete against with each other.

The concept of the sports league matches brings in a new wave of entertainment, enthusiasm, competition and professionalism in promoting the particular sport and sports people's persona and economic wellbeing. It plays a big role in promoting businesses around sports industry. It has emerged as a big business among the sports fraternity.

3.2 History

International Association for The Sports Information (IASI) with the goal of standardizing and archiving the world's sports libraries (International Association for Sports Information, 2008), was founded in 1960. The International Association on Computer Science in Sport (IACSS) to better improve the co-operation amongst The international researchers who are interested in applying Computer Science techniques and technologies to sport-related challenges (International Association on Computer Science in Sport, 2008).was founded in 1997

The IASI is a worldwide network of the sport librarians, experts, and document repositories. The Association's information dissemination comes in the form of a tri-annual newsletter and an organized World Congress every four years. The IACSS focuses on the disseminating and the research of their members with the periodic newsletters and the journals and biannual conferences.

3.3 Sports League Types

3.3.1 Hockey India League (HIL)

Hockey India league is second a successful and popular leagues after Indian Premier Leagues in India as well as in the world it's basically stick and ball game played between two teams each contains eleven players each it is a nineteen minute game with two half of forty five minute.

Hockey India League(HIL) is a Hockey Competition initiated by Hockey India was started from 2013 consisting of 5 franchise, when Hockey players from different countries can participate since then HIL has become very popular throughout the World-Wide, a new Team Kalinga Lancers joined further this will increase the number of Franchise from 5 to 6.

3.3.2 Indian Primer League (IPL)

Indian Premier League (IPL) is a Twenty-twenty cricket competition commenced by the (BCCI) Board of Control for Cricket in India, located in Mumbai. It was started in year 2008 with 8 franchises, where players from different cricketing countries can play. IPL become much popular throughout the world. It was announced On 21 March 2010 that two new teams Pune Warriors and Kochi Tuskers will be added for IPL 4th edition and the number of franchises will increase from 8 to 10 , but later both the new teams left the league and now total 8 teams are in Indian Primer League.

4.3.3 Indian Badminton League (IBL)

The Indian Badminton League started from 14 august 2013 to 31 august 2013 .In India the Governing body for Badminton is Badminton Association Of India (BAI) and is located in Lucknow, BAI as association registered under the Indian Olympic Association (IOC), Badminton World Federation (BWF), and Badminton Asia Confederation (BAC), the National level tournament held by (BAI) since 1936 and it was formed in 1934.

BAI organizes International tournaments in India like Syed Modi India Grand Prix, India Open Super Series, and the national championships in nearly every age categories, and from all the above tournaments selects the team and conducts the training camp for international tournaments such as World Championship, Olympic Games and the training of umpires, technical officials, path-breaking initiatives like (IBL) to promote the sport.

3.4.4 Pro Kabaddi League (PKL)

The Pro Kabaddi League (PKL) started earlier in 2014 and it is a professional kabaddi league in India, based on the format of Hockey India League (HIL) and the Indian Premier League (IPL). The first edition was started on 26 July 2014 with 8 teams from different part of country consisting of players from the entire world. The Pro Kabaddi League (PKL) is the first significant initiative of Marshal Sports.

Pro Kabaddi League (PKL) took our truly indigenous sport of Kabaddi to a new level of professionalism, which will benefit all the players, as all other league games stakeholders involved in the ecosystem of the Kabaddi, most of all, who was not known before will become the new role models for the youth of India. Pro Kabaddi League is eight-team league with and games will be played on a caravan format and each team will play with each team twice in month of July and August, 2014.

4. Experimental Set Up and Test Data

The concept of clustering has been considered in order to classify HIL goals statistics data into appropriate clusters using k-mean and fuzzy algorithm [11]. A database has been constructed and the whole records comprises of 21 players which are collected from HIL website. The dataset consist of several attributes like player-id, player name, team name, matches, goals, average and membership function which are clearly shown in figure 3.

4.1 Membership Function

The membership function is generated from the database corresponding to each record into values which lies in the range of 0 to 1 taking goals/matches as parameter. The membership function and its corresponding graph are shown in Figure 2. and the graph of player id and the membership value are shown in the Figure 4 and 5.

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

Figure 2: Membership Functions

where x = goals/match,
a = minimum of x,
b = median of x and
c = maximum of x.

Centroid of kth cluster:-

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

ID	PLAYER	TEAM	MATCHES	GOALS	AVERAGE	MEMBERSHIP FUNCTION	C1(0.2)	C2(0.4)	CLUSTER
1	SANDEEP SINGH	PUNJAB	16	15	0.93	0	0.2	0.4	1
2	V.R RAGUNATH	UP	17	13	0.76	0.153846153	0.046153847	0.246153847	1
3	MANDEEP SINGH	RANCHI	15	11	0.73	0.307692307	0.107692307	0.092307693	2
4	RUPINDER PAL SINGH	DELHI	17	10	0.58	0.714285714	0.514285714	0.314285714	2
5	ASHLEY JACKSON	RANCHI	15	7	0.46	0.615384615	0.415384615	0.215384615	2
6	LLOYD NORRIS	DELHI	17	7	0.41	0.615384615	0.415384615	0.215384615	2
7	DHARAMVIR SINGH	PUNJAB	18	6	0.33	0.692307692	0.492307692	0.292307692	2
6	SIMON CHILD	DELHI	17	5	0.29	0.769230769	0.569230769	0.369230769	2
7	JUSTIN REID ROSS	RANCHI	15	5	0.33	0.769230769	0.569230769	0.369230769	2
8	JAMIE DWYER	PUNJAB	18	5	0.27	0.769230769	0.569230769	0.369230769	2
9	MALAK SINGH	PUNJAB	16	5	0.31	0.769230769	0.569230769	0.369230769	2
10	JASON WILSON	DELHI	11	4	0.36	0.846153846	0.646153846	0.446153846	2
11	S V SUNIL	PUNJAB	16	4	0.25	0.846153846	0.646153846	0.446153846	2
12	CHRISTOPHER CIRIELLO	PUNJAB	18	4	0.22	0.846153846	0.646153846	0.446153846	2
13	AKASHDEEP SINGH	DELHI	18	3	0.16	0.923076923	0.723076923	0.523076923	2
14	GLENN TURNER	MUMBAI	13	3	0.23	0.923076923	0.723076923	0.523076923	2
15	ANDRES MIR	DELHI	18	3	0.16	0.923076923	0.723076923	0.523076923	2
16	DANISH MUJTABA	DELHI	16	2	0.12	1	0.8	0.6	2
17	AUSTIN SMITH	RANCHI	13	2	0.15	1	0.8	0.6	2
18	MANPREET SINGH	RANCHI	17	2	0.11	1	0.8	0.6	2
19	AFFAN YOUSOUF	PUNJAB	4	2	0.5	1	0.8	0.6	2

Figure 3: HIL Data Set

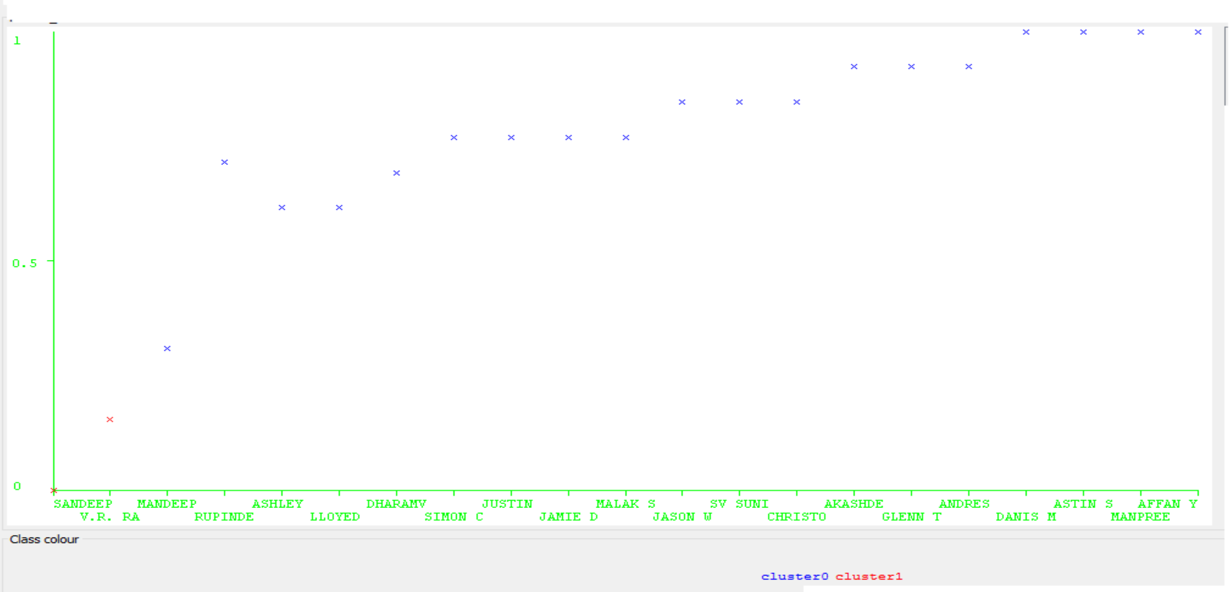


Figure 4: Player Vs Membership Function (K-Means Algorithm)

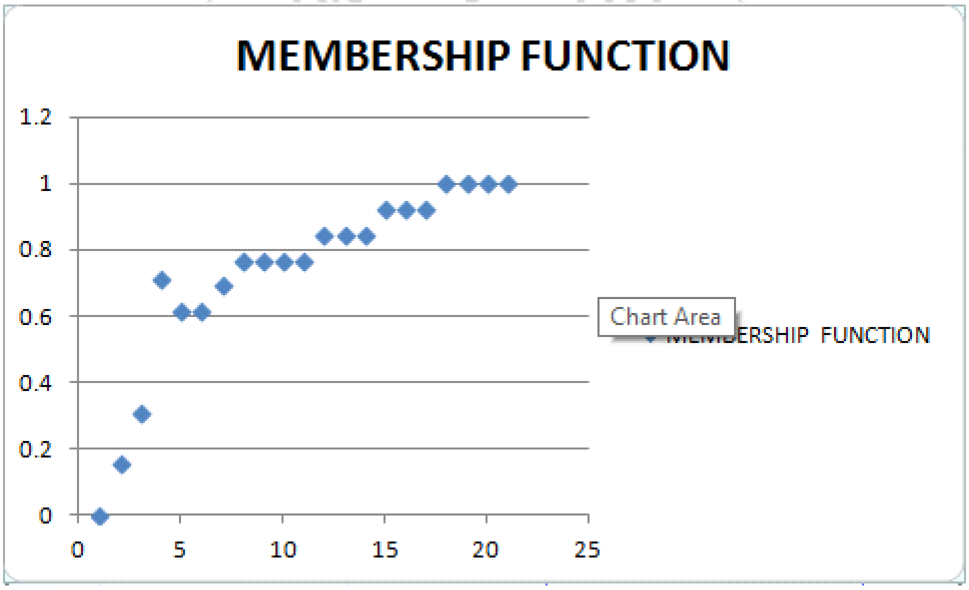


Figure 5: Player Vs Membership Function (Fuzzy Algorithm)

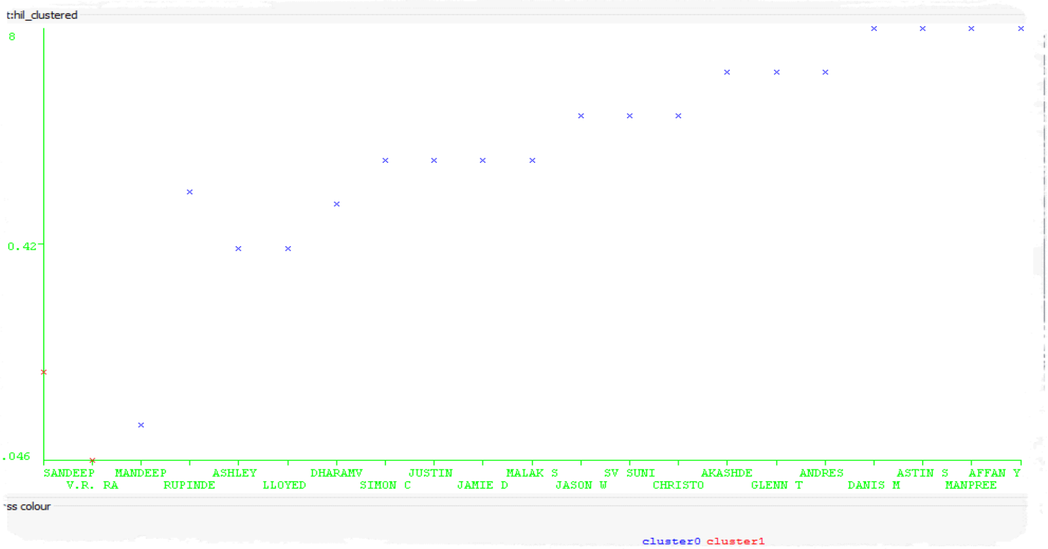


Figure 6: Player Vs Cluster (K-Means Algorithm)

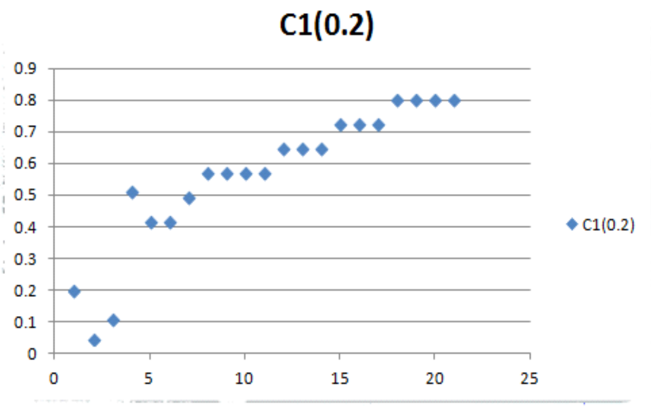


Figure 7: Player Vs Cluster (Fuzzy Algorithm)

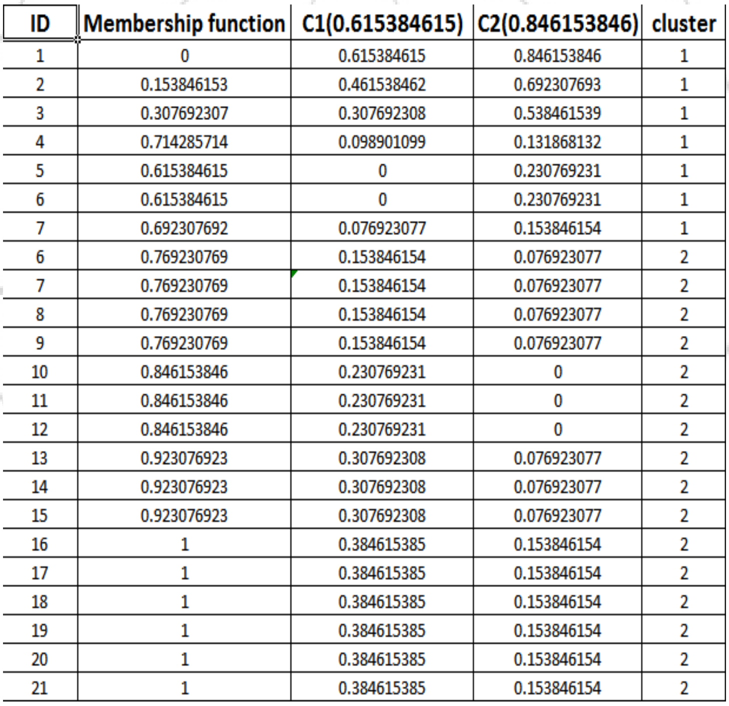
5. Initialization Problem

The K-means is the simplest and most commonly used algorithm employing a squared error criterion. It starts with random initial centroids and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centroids until a convergence criterion is met after some number of iterations. The K-means algorithm is popular because it is easy to implement, and its time complexity is O (n), where n is the number of patterns.

A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial centroids are not properly chosen

6. Solution to this Problem

To solve this problem an alternate initialization technique was used.
"MacQueen" initialization technique was used which consisted of the following steps [5]
Step1: Choose k objects at random and use them as initial centroid.
Step2: Assign each object to the cluster with the nearest centroid.
Step3: After each assignment, recalculate the centroid.



ID	Membership function	C1(0.615384615)	C2(0.846153846)	cluster
1	0	0.615384615	0.846153846	1
2	0.153846153	0.461538462	0.692307693	1
3	0.307692307	0.307692308	0.538461539	1
4	0.714285714	0.098901099	0.131868132	1
5	0.615384615	0	0.230769231	1
6	0.615384615	0	0.230769231	1
7	0.692307692	0.076923077	0.153846154	1
6	0.769230769	0.153846154	0.076923077	2
7	0.769230769	0.153846154	0.076923077	2
8	0.769230769	0.153846154	0.076923077	2
9	0.769230769	0.153846154	0.076923077	2
10	0.846153846	0.230769231	0	2
11	0.846153846	0.230769231	0	2
12	0.846153846	0.230769231	0	2
13	0.923076923	0.307692308	0.076923077	2
14	0.923076923	0.307692308	0.076923077	2
15	0.923076923	0.307692308	0.076923077	2
16	1	0.384615385	0.153846154	2
17	1	0.384615385	0.153846154	2
18	1	0.384615385	0.153846154	2
19	1	0.384615385	0.153846154	2
20	1	0.384615385	0.153846154	2
21	1	0.384615385	0.153846154	2

Figure 8: MacQueen DataSet1

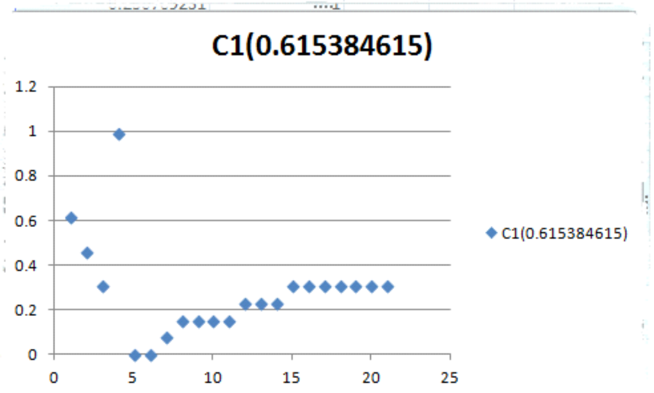
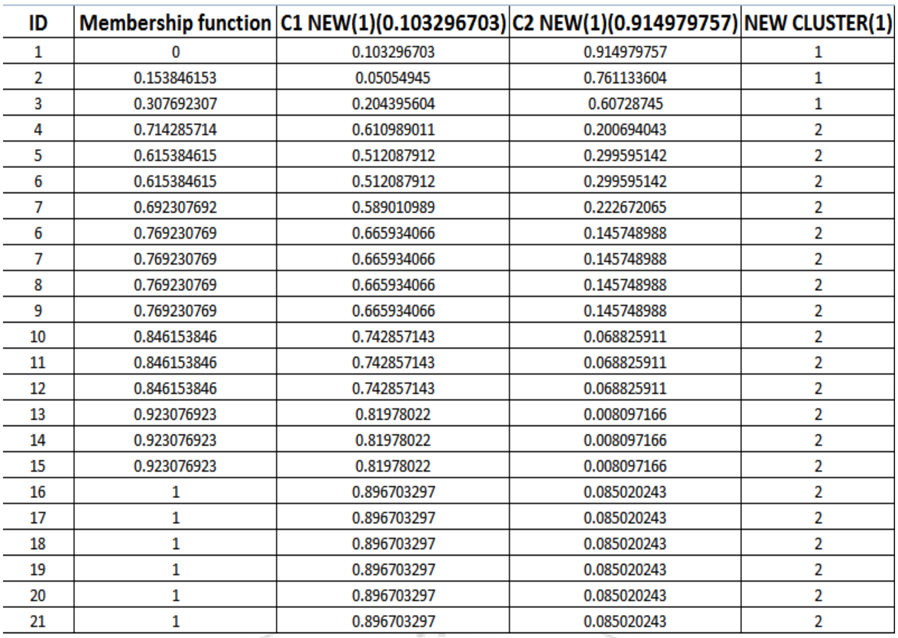


Figure 9: Player Vs Cluster (MacQueen 1)



ID	Membership function	C1 NEW(1)(0.103296703)	C2 NEW(1)(0.914979757)	NEW CLUSTER(1)
1	0	0.103296703	0.914979757	1
2	0.153846153	0.05054945	0.761133604	1
3	0.307692307	0.204395604	0.60728745	1
4	0.714285714	0.610989011	0.200694043	2
5	0.615384615	0.512087912	0.299595142	2
6	0.615384615	0.512087912	0.299595142	2
7	0.692307692	0.589010989	0.222672065	2
6	0.769230769	0.665934066	0.145748988	2
7	0.769230769	0.665934066	0.145748988	2
8	0.769230769	0.665934066	0.145748988	2
9	0.769230769	0.665934066	0.145748988	2
10	0.846153846	0.742857143	0.068825911	2
11	0.846153846	0.742857143	0.068825911	2
12	0.846153846	0.742857143	0.068825911	2
13	0.923076923	0.81978022	0.008097166	2
14	0.923076923	0.81978022	0.008097166	2
15	0.923076923	0.81978022	0.008097166	2
16	1	0.896703297	0.085020243	2
17	1	0.896703297	0.085020243	2
18	1	0.896703297	0.085020243	2
19	1	0.896703297	0.085020243	2
20	1	0.896703297	0.085020243	2
21	1	0.896703297	0.085020243	2

Figure 10: MacQueen Dataset (2)

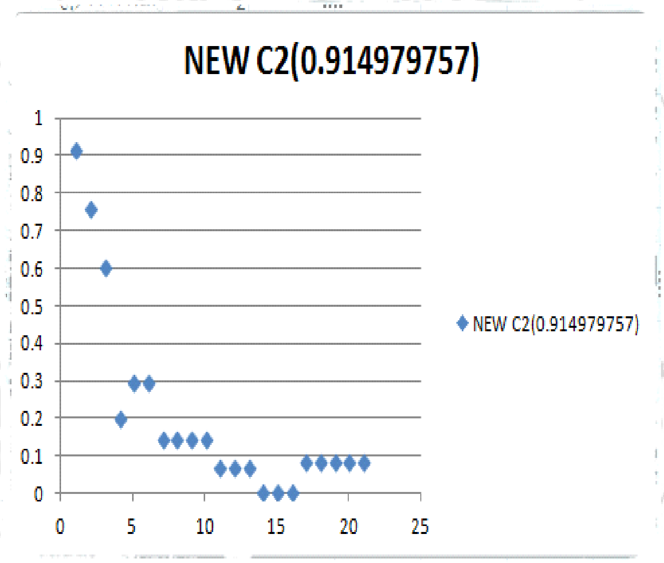
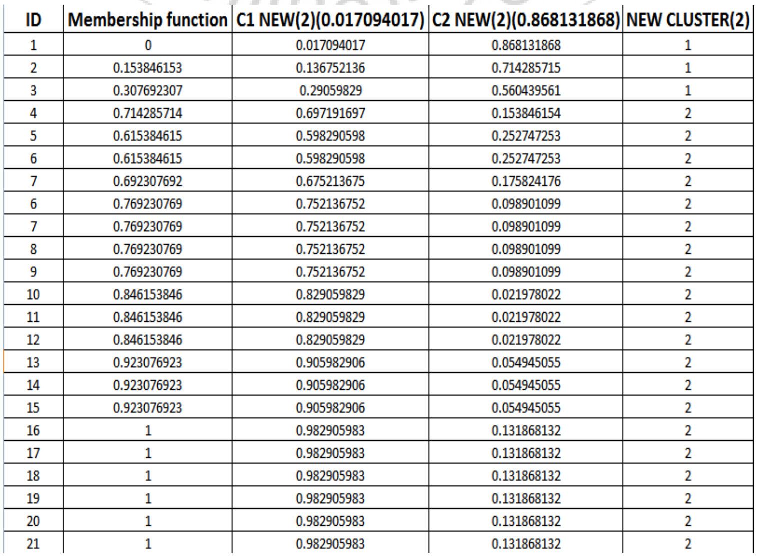


Figure 11: MacQueen Dataset (2)



ID	Membership function	C1 NEW(2)(0.017094017)	C2 NEW(2)(0.868131868)	NEW CLUSTER(2)
1	0	0.017094017	0.868131868	1
2	0.153846153	0.136752136	0.714285715	1
3	0.307692307	0.29059829	0.560439561	1
4	0.714285714	0.697191697	0.153846154	2
5	0.615384615	0.598290598	0.252747253	2
6	0.615384615	0.598290598	0.252747253	2
7	0.692307692	0.675213675	0.175824176	2
6	0.769230769	0.752136752	0.098901099	2
7	0.769230769	0.752136752	0.098901099	2
8	0.769230769	0.752136752	0.098901099	2
9	0.769230769	0.752136752	0.098901099	2
10	0.846153846	0.829059829	0.021978022	2
11	0.846153846	0.829059829	0.021978022	2
12	0.846153846	0.829059829	0.021978022	2
13	0.923076923	0.905982906	0.054945055	2
14	0.923076923	0.905982906	0.054945055	2
15	0.923076923	0.905982906	0.054945055	2
16	1	0.982905983	0.131868132	2
17	1	0.982905983	0.131868132	2
18	1	0.982905983	0.131868132	2
19	1	0.982905983	0.131868132	2
20	1	0.982905983	0.131868132	2
21	1	0.982905983	0.131868132	2

Figure 12: MacQueen Dataset (3)

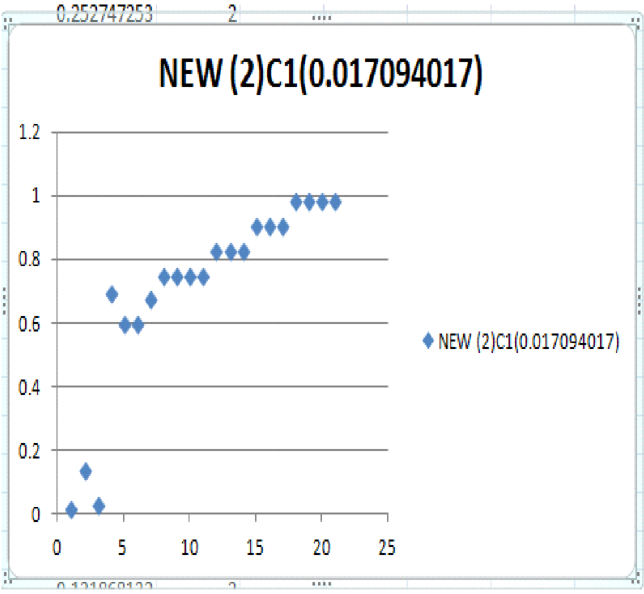


Figure 13: MacQueen Dataset (3)

7. Conclusion and Future Directions

Data Clustering plays a major role in grouping the similar type of data into a specific cluster. Cluster targets at selecting groups of same objects and, then helps to find out distribution of patterns and interesting correlations in huge data sets. K-Means clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by using different attributes. Fuzzy clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships.

In this paper, K-Means and Fuzzy clustering has been adopted using HIL database to detect two clusters on the HIL 2014 goals statistics dataset. The records in the database are partitioned in a way such that same types of records are grouped in the similar cluster. N-clusters have been detected from HIL goals statistics dataset. WEKA has been used for the definition the membership function, and detecting the several clusters. Also the initialization problem that was faced in K-Means was solved in this paper.
The future research work of this article lies on the fact that all version of HIL dataset will be taken into consideration to develop an algorithm which detects n-clusters to generalize the K-Means and Fuzzy clustering techniques and comparison of all version of HIL 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, seconded, Morgan Kaufmann, California, 2005.

[3] 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.

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

[5] MACQUEEN, J. "Some methods for classification and analysis of multivariate observations", 1967.

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

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

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

[9] Shehroz and Ahmad, "Cluster center initiation algorithm for k-means clustering" , 2004.

[10] www.hil.com

[11] 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.