Application of Fuzzy K-Means (FKM) Algorithms in Identifying Better Clusters of Few Drugs from Drugbank Database

Naga Madhavi Latha Kakarla, G.Rama Mohan Babu

Abstract—Thefuzzification of the cluster configuration is refereed as Fuzzy K-Means (FKM) where the algorithm generates limited homogeneous clusters. The data points are assigned respective clusters in accordance to the membership degrees within interval [0,1]. Several variations of FKM algorithm were applied in identifying better clusters of few drugs data set derived from DrugBank database as possible GSK-3 beta inhibitors defined against diabetes. Better clusters were evaluated based on cluster balance and membership degree plots. With k=3, observation of cluster balance and membership degree plots revealed that FKM with entropy is the best method of choice with equal assignment of objects and no ambiguous assignments. The membership degree plot resulted in a good fuzzy clustering result where only 3 points appeared between membership degrees of 0.6 to 0.8

Keywords— fuzzy k-means, FKM, clustering, GSK-3 beta, cluster balance, membership degree

1. INTRODUCTION

Generalization of partition clustering is done by fuzzy clustering as it allows partial classification of an object into clusters > 1 [1]. In fuzzy clustering, the membership is spread among all clusters. Fuzzy k-means (FKM) clustering algorithm has received attention by several researchers [2] [3]. The FKM algorithm generates limited number of homogeneous clusters. Depending on the membership values which lay between the range (0 and 1) the objects are assigned to their respective clusters. This is referred as a fuzzification. The constraint for the memberships is that they should be non negative. The sum of membership values for an object should be one. Here, the constraints of memberships are same. These constraints appear as probabilities of objects to belong to particular group [4]. The fuzzy clustering has a merit of not forcing all objects into a specific cluster [5]. The most difficult tasks in cluster analysis isthe selection of appropriate number of clusters. In fuzzy clustering, the coefficients are used along with silhouette scores. The fuzziness in a solution is measured by Dunn's partition coefficient which processes how close the fuzzy solution is to the corresponding hard solution. This is formed by classifying each object into the cluster which has the largest membership value [6].

The classicalk-means algorithm is capable of discovering hard clusters where an observation belongs to only one cluster so that the objects of the same cluster are similar and different clusters are dissimilar, whereas the Fuzzy K-Means is more statistically inclined method which results in soft clusters where a particular observation has probability to appear in more than one cluster. The k-means algorithm is widely presented in many cluster dataset problems and gained more efficiency, however, a variety of modifications to classical k-means have been proposed and developed. Among them, FKM is more popular, originally proposed by Ruspini [7] and modified by Bezdek [8]. The k-means algorithm is anexample for hard clustering, whereas FKM results in soft clustering.

In this paper, we report the several variations of FKM algorithm,were applied, in identifying better clusters of few drugs data set derived from DrugBank database as possible GSK-3 beta inhibitors defined against diabetes.

2. MATERIALS AND METHODS

2.1 DATASET

The top 13 drugs reported as better GSK-3 beta inhibitors in our previous study were selected as dataset.

Revised Manuscript Received on April 12, 2019.

Naga MadhaviLathaKakarla, Research Scholar, Department of Computer Science and Engineering, University College of Engineering & Technology, AcharyaNagarjuna University, Nagarjuna Nagar, GUNTUR - 522510, Andhra Pradesh, India. (E-mail: kakarla.nml@gmail.com)

G.Rama Mohan Babu, Professor, Department of Information Technology, R.V.R. & J.C College of Engineering, Chowdavaram GUNTUR- 522019, Andhra Pradesh, India. (E-mail: rmbgatram@gmail.com)

Table 1: Dataset considered for analysis. Observations are parameters calculated by docking program, Molegro using default values.

Ligand	Mol Dock Score (kcal/mol)	Rerank Score (kcal/mol)	Inter action energy (kcal/mol)	Protein energy (kcal/mol)	Internal energy (kcal/mol)	No. of Tor sions	HBond energy	Mole cular Weight	LE1	LE3
DB00183	-156.515	-88.3902	-194.357	-194.357	37.8419	23	-11.5715	767.891	-2.89843	-1.63685
DB01076	-155.873	-107.164	-193.494	-193.494	37.6204	12	-7.99997	557.632	-3.80179	-2.61376
DB06590	-186.456	-89.6677	-193.677	-193.677	7.22099	11	-13.5466	684.685	-4.33618	-2.0853
DB06441	-178.802	-129.546	-188.25	-188.25	9.44806	16	-19.1445	776.359	-4.06368	-2.94422
DB06695	-172.986	-84.6672	-185.692	-185.692	12.7063	18	-4.59018	627.733	-3.76056	-1.84059
DB00503	-155.481	-86.1033	-193.886	-193.886	38.4045	21	-7.74278	720.944	-3.10963	-1.72207
DB00966	-173.267	-65.935	-173.405	-173.405	0.13778	7	-4.27836	514.617	-4.44275	-1.69064
DB08822	-172.861	-74.5849	-157.617	-157.617	-15.2446	10	-6.33187	568.534	-4.11575	-1.77583
DB08909	-165.266	-85.1091	-176.113	-176.113	10.8475	20	-3.03005	530.651	-4.23759	-2.18228
DB11581	-202.148	-111.509	-234.88	-234.88	32.7314	13	-0.78332	868.439	-3.31391	-1.82802
DB09050	-185.412	-72.5561	-184.811	-184.811	-0.60118	14	-15.2516	666.69	-4.12026	-1.61236
DB09065	-181.14	-92.5787	-211.613	-211.613	30.4724	23	-7.89876	776.023	-3.35445	-1.71442
DB09230	-157.455	-96.1705	-166.629	-166.629	9.17463	11	-4.14597	582.646	-3.66173	-2.23652

LE1: Ligand Efficiency 1, dock score divided by heavy atoms count
LE2: Ligand Efficiency 2, re-rank score divided by heavy atoms count

2.2 Fuzzy k-means

The algorithms used in fuzzy k-means procedure are:

- FKM algorithm
- Fuzzy k-means with regularization of entropy
- Fuzzy k-means with regularization of entropy and noise cluster
- Gustafson and Kessel - like fuzzy k-means
- Gustafson and Kessel - like fuzzy k-means with regularization of entropy
- Gustafson and Kessel - like fuzzy k-means with regularization of entropy and noise cluster

Fuzziness parameter (m)

The fuzziness parameter, *m* (membership value) depends on the dataset and variations of objects within the dataset. The value of *m* should not be less than 1, hence work was evaluated to find appropriate value of *m* by varying its values. The resultant cluster plots were analyzed for better clustering solutions. It was reported by Bezdek, (1981) [9] that when m goes to infinity, values of object attains 1/K, which suggests that for a given dataset, at a particular value of *m* and above which the membership values obtained from fuzzy c-means equals to 1/K.

3. RESULTS AND DISCUSSION

3.1 FKM algorithm:

The dataset comprising 13 data points from DrugBank database screened against GSK-3 beta as positive drugs dataset was subjected to fkm algorithm and evidenced that the program is able to cluster 3 sets. For all the methods studied, membership coefficients, visual assessment of cluster tendency for drugs dataset were recorded (data not shown).

The quality and acceptability of the solution is inspected by VIFCR. The cluster balance Figure-1 shows nearly equal assignment of objects with limited number of ambiguous assignments. Further, the conformation of fkm fuzzy clustering result is done by max membership degree plot where almost all points appeared between membership degrees 0.4 to 0.8, which represents not a better plot (Figure 2).

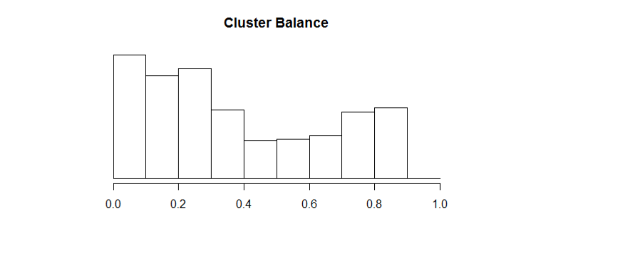

Figure 1: Cluster balance for fkm fuzzy clustering results on dataset.

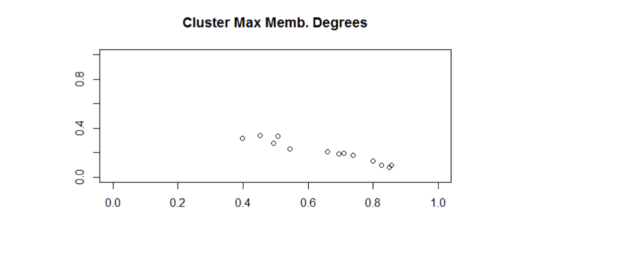

Figure 2: The highest membership degrees for fkm fuzzy clustering results on drugs dataset.

3.2 Fuzzy k-means with regularization of entropy

This is an algorithm that is involved in performing the fuzzy k-means clustering analysis with regularization of entropy. The regularization of entropy avoided using the artificial fuzziness parameter *m*. Instead of *m*, a degree of fuzzy entropy *ent*is provided [10]. It is the term of statistical physics used to denote the physical quantity temperature. A unique feature of fuzzy k-means with regularization of entropy is that the prototypes are generated as mean of weights where membership values are nothing but weights. In fuzzy - means weights are equal to membership values at the power of *m*.

The quality and acceptability of the solution is inspected by VIFCR. The cluster balance plot given in figure-3 showed almost equal assignment of objects with no ambiguous assignments. Further, a very good fuzzy clustering result is shown by maximum membership degree plot where only 3 points appeared between membership degrees of 0.6 to 0.8 (Figure 4).

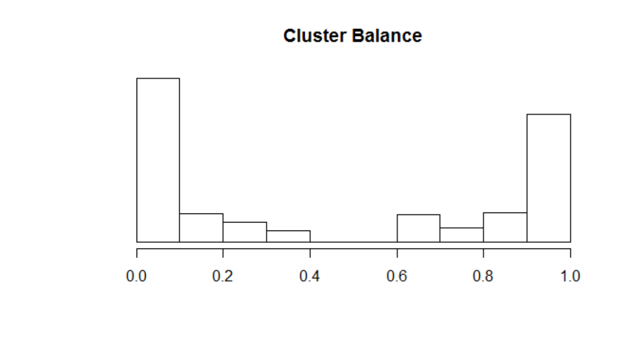

Figure 3: Cluster balance for fkment fuzzy clustering results on dataset.

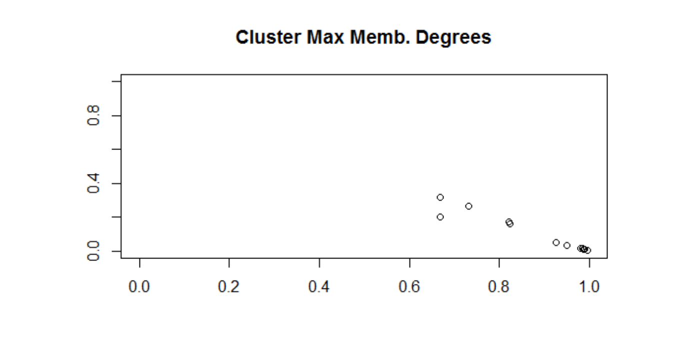

Figure 4: The highest membership degrees for fkment fuzzy clustering results on drugs dataset.

3.3 Fuzzy k-means with regularization of entropy and noise cluster

This is an algorithm that is involved in performing the fuzzy k-means clustering with regularization of entropy and noise cluster. The regularization of entropy avoided using the artificial fuzziness parameter *m*. Noise cluster is one to which objects identified at boundaries are assigned. The objects in this cluster possess high membership degrees [11]. The k standard clusters maintain homogeneity, whereas the noise cluster is not homogenous. It is formed by all the objects that are at boundaries and are not homogenous.

The cluster balance plot given in figure-5 showed exact equal assignment of objects with no ambiguous assignments as the program nullified noise. A fuzzy clustering result is confirmed by maximum membership degree plot where only 1 point appeared at zero membership degree (Figure 6).

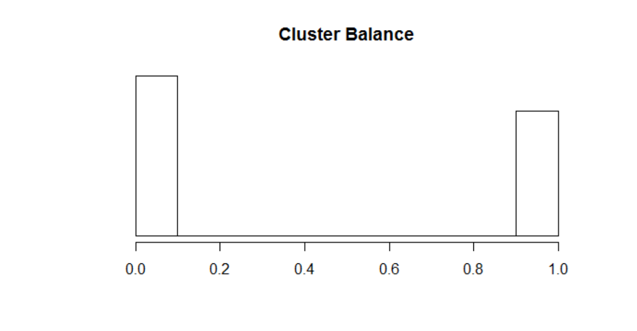

Figure 5: Cluster balance for fkmentnoise fuzzy clustering results on dataset

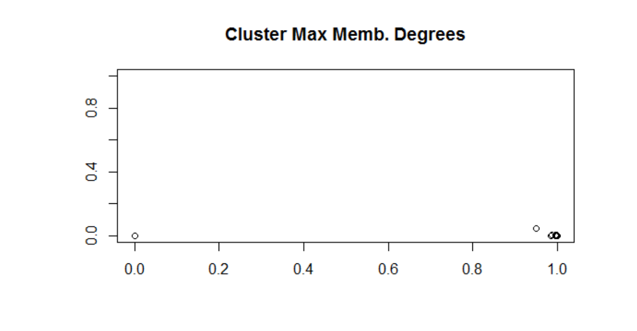

Figure 6: The highest membership degrees for fkmentnoise fuzzy clustering results on drugs dataset

3.4 Gustafson and Kessel - like fuzzy k-means

The program executes the algorithm, Gustafson and Kessel - like fuzzy k-means clustering, and is it can identify non-spherical clusters [12]. The cluster balance plot given in figure-7 showed ambiguous assignments and the fuzzy clustering result confirmed by maximum membership degree plot observed all points appeared between 0.4 to 0.8 membership degrees.

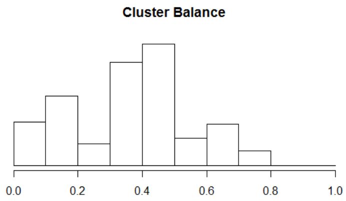	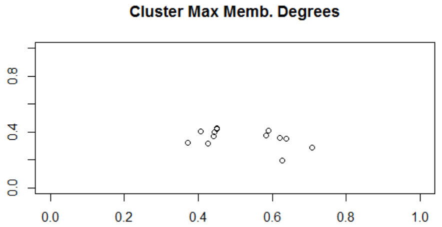
Figure 7: Cluster balance for fkmgk fuzzy clustering results on dataset	**Figure 8: The highest membership degrees for fkmgk fuzzy clustering results on drugs dataset**

3.5 Gustafson and Kessel - like fuzzy k-means with regularization of entropy

The program executes the algorithm, Gustafson and Kessel - like fuzzy k-means clustering, with regularization of entropy [13]. The regularization of entropy makes us convenient to avoid the use of artificial fuzziness parameter *m*. Running the clustering program is done by using standardized data if standardization is set to *stand=1*.

The cluster balance plot given in figure-9 is similar to fkmgk method where it showed ambiguous assignments and the fuzzy clustering result confirmed by maximum membership degree plot (Figure 10) observed all points appeared between 0.4 to 0.8 membership degree.

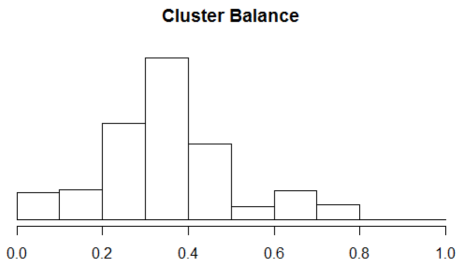	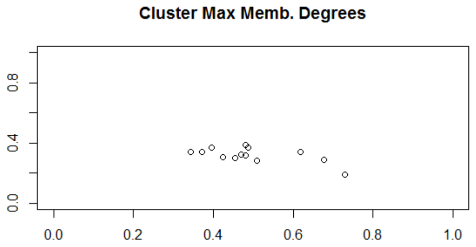
Figure 9: Cluster balance for fkmgkent fuzzy clustering results on dataset	**Figure 10: The highest membership degrees for fkmgkent fuzzy clustering results on drugs dataset**

3.6 Gustafson and Kessel - like fuzzy k-means with regularization of entropy and noise cluster

The program executes the algorithm, Gustafson and Kessel - like fuzzy k-means clustering, with regularization of entropy and noise cluster which is different from fuzzy k-means and was unable to identify non-spherical clusters. The cluster balance plot given in figure-11 is similar to fkmgk and entropy method where it showed ambiguous assignments and the fuzzy clustering result confirmed by maximum membership degree plot (Figure 12) observed all points appeared between 0.4 to 0.8 membership degrees.

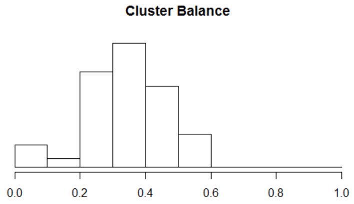	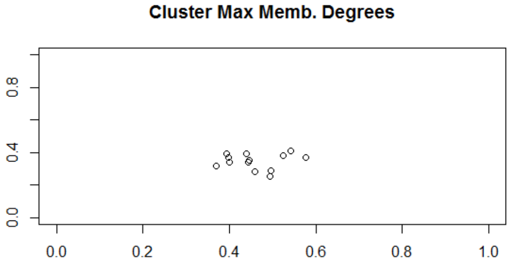
Figure 11: Cluster balance for fkmgkentnoise fuzzy clustering results on dataset	**Figure 12: The highest membership degrees for fkmgkentnoise fuzzy clustering results on drugs dataset**

The number of clusters formed by each method was given for comparison in Table 2. From the table it was observed that each method resulted in varied number of clusters in each group with k=3. Hence, observation of cluster balance and membership degree plots revealed that fkment is the best method of choice with better separation of clusters with equal assignment of objects and no ambiguous assignments. The membership degree plot resulted in a good fuzzy clustering result where only 3 points appeared between membership degrees of 0.6 to 0.8.

Table 2: Number of clusters formed by each method.

Method	Cluster1	Cluster2	Cluster3
Fkm	4	5	4
Fkment	3	6	4
Fkmentnoise	4	4	5
Fkmgk	6	2	5
Fkmgkent	3	4	6
Fkmgkentnoise	5	4	4

Validation Studies:

Several validation studies has been proposed in clustering datasets, however, not all validation procedures are required in one aspect. Hence, following few validation statistics are presented in Table 3. They are:

Partition Coefficient index (PC): Maximization of index value resulted in the achievement of optimal number of cluster *k.*

Partition Entropy index (PE): Minimization of index value resulted in the achievement of optimal number of cluster *k.*

Modified partition coefficient index (MPC): Maximization of index value resulted in the achievement of optimal number of cluster *k.*

Silhouette index (SIL): Maximization of index value resulted in the achievement of optimal number of cluster *k.*

Fuzzy silhouette index (SIL.F): Maximization of index value resulted in the achievement of optimal number of cluster *k.*

Xie and Beni index (XB): Minimization of index value resulted in the achievement of optimal number of cluster *k.*

Table 3:Validation index values of fkm algorithms

Algorithm	PC	PE	MPC	SIL	SIL.F	XB
Limit Value (Max/Min)	**Max**	**Min**	**Max**	**Max**	**Max**	**Min**
Fkm	0.5284	0.8187	0.2927	0.3541	0.4296	0.4543
fkm.ent	**0.8228**	**0.3145**	**0.7342**	**0.3975**	**0.4334**	**0.5636**
fkm.ent.noise	0.8318	0.0252	0.7477	0.3029	0.3090	0.3091
Fkm.gk	0.436	0.9043	0.1548	-0.262	-0.277	-
Fkm.gk.ent	0.4028	0.9794	0.1043	-0.175	-0.197	-
Fkm.gk.ent.noise	0.5756	0.6118	0.1513	0.0904	0.1314	-

From table 3, it is evidenced that the fuzzy k-means algorithm resulted in parameter values within the limits. Overall fkm.ent method was able to produce better solution. Hence, the drugs which appeared in each cluster represent significant relationships and further analysis on their biochemical and enzymatic properties would reveal functional activities.

4. CONCLUSION

FKM as a method of choice to cluster was reported on few drugs that are computationally evaluated to inhibit GSK-3 beta protein *in vitro*. Nearly six variations of FKM algorithms were studied and analysis revealed that fkm with entropy is the best method of choice with better separation of clusters with equal assignment of objects and no ambiguous assignments based on cluster balance and membership degree plots. Therefore it is anticipated that FKM with entropy method performed better than other alternatives for this particular dataset.

REFERENCES

1. Neumaier A. Clouds, fuzzy sets, and probability intervals. Reliable computing. 2004 Aug 1;10(4):249-72.
2. J.C. Bezdek, (1974). Numerical taxonomy with fuzzy sets, Journal of Mathematical Biology 1 (1974) 57–71.
3. E.A. Ruspini, A new approach to clustering, Information and Control 15 (1969) 22–32.
4. Zadeh LA. Fuzzy algorithms.Information and control. 1968 Feb 1;12(2):94-102.
5. Zadeh LA. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh. World Scientific; 1996.
6. Rousseeuw PJ, Kaufman L, Trauwaert E. Fuzzy clustering using scatter matrices. Computational Statistics & Data Analysis. 1996 Nov 15;23(1):135-51.
7. Enrique H Ruspini. A new approach to clustering.Information and control, 15(1):22–32, 1969
8. James C Bezdek.A convergence theorem for the fuzzy isodata clustering algorithms. IEEE Transactions on Pattern Analysis & Machine Intelligence, (1):1–8, 1980.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

9. J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.
10. Li R., Mukaidono M., 1995. A maximum entropy approach to fuzzy clustering. Proceedings of the Fourth IEEE Conference on Fuzzy Systems (FUZZ-IEEE/IFES ’95), pp. 2227-2232.
11. Dave’ R.N., 1991. Characterization and detection of noise in clustering. Pattern Recognition Letters, 12, 657-664..
12. Gustafson E.E., Kessel W.C., 1978. Fuzzy clustering with a fuzzy covariance matrix. Proceedings of the IEEE Conference on Decision and Control, pp. 761-766.
13. Ferraro M.B., Giordani P., 2013. A new fuzzy clustering algorithm with entropy regularization.Proceedings of the meeting on Classification and Data Analysis (CLADAG).

AUTHORS’ PROFILES

Kakarla Naga MadhaviLatha is aresearch scholar in the Department of Computer Science and Engineering in AcharyaNagarjuna University, Guntur India. She is working as Assistant Professor in Sir C R Reddy College of Engineering, ELURU. India. She has 12 years of teaching experience in Computer Science and Engineering, and vast research experience. Her areas of research are data mining and bioinformatics.

G. Rama Mohan Babu is working as a Professor in the Department of Information Technology in R.V.R. & J.C College of Engineering, Chowdavaram, Guntur, India. He received his Ph.D in CSE fromAcharyaNagarjuna University, Guntur, India. He has 18 years of experience in teaching and research in Computer Science & Engineering. His areas of interest are image processing, pattern recognition, semantic web, bioinformatics and data mining.