Fuzzy Clustering algorithms and their applications to chemical datasets

Jehan Zeb Shah, Ph.D. student , Email: zeb@scientist.com
PM Dr. Naomie bt Salim, Supervisor, Email: naomie@fsksm.utm.my

Abstract:

In this work the importance of fuzzy based clustering methods is highlighted and their applications in the field of chemoinformatics, and issues involved are reviewed. The various methods and approaches of fuzzy clustering are outlined. The issue of number of valid clusters in a dataset is also discussed. The hyper dimensional chemical datasets are traditionally been treated only with the help of conventional clustering methods like hierarchical and non-hierarchical methods. In this paper we look into the issue of clustering these chemical datasets with fuzzy paradigms. In this paper a number of fuzzy clustering approaches like fuzzy c-mean, Gustafson and Kessel , Gath and Geva, fuzzy c-varieties, adaptive fuzzy , fuzzy based c-shell algorithms and some other aspects of fuzzy clustering are discussed.

Key words
fuzzy c-mean, clustering, chemoinformatics, neural network.

1. Introduction

Fast and cost effective drug discovery methods are the main objective of the new field of chemoinformatics. The drug discovery is a complex and costly process in its essence. Previously, the main bottlenecks in drug discovery were the time and costs of making (or finding) and testing new chemical entities (NCE). The average cost of creating a NCE in a major pharmaceutical company was estimated at around 7,500/compound [2]. In order to reduce costs, pharmaceutical companies have had to find new technologies to replace the old "hand-crafted" synthesis and testing NCE approaches.

By 2000, many solution- and solid- phase combinatorial chemistry strategies were well developed [3]. So, millions of new compounds can be created by these CC based technologies. But the problem is that these procedures are completely failed to yield many drug candidates (compounds that can be turned into drugs are also called drug like compounds). So, tracing the reasons for these disappointing results, it is believed that enhancing the chemical diversity of compound libraries would enhance the drug discovery process. In order to arrive at a chemical library of great chemical diversity, a number of structural processing technologies like structural descriptor computations; structural similarity algorithms, library enumerations, diversified compound selections, and classification and clustering algorithms have been developed. The clustering of chemical data sets forms the real estate of this paper.

The main objective of Clustering (or cluster analysis) is to organize a collection of data items into some meaningful clusters, so that items within a cluster are more similar to each other than they are to items in the other clusters. This notion of similarity and dissimilarity may be based on the purpose of the study or domain specific knowledge. But one thing must be kept in mind that from clustering we always mean unsupervised classification of data, a course where is no teacher to provide any guidance of the path. In other words, there is no pre notion about the groups and their number (may or may not be) present in the data set.

Most traditional cluster analysis methods are crisp (or hard) partitioning, in which every given object is strictly classified into a certain group. So, the boundaries defined for the objects (data elements) are very sharp and so they go to only and only one cluster. However, the features or attributes of the objects in practice are not sharp and they may be having some tendency to be the part of some class to some extent. Fortunately, the fuzzy sets theory proposed by Zane [1] provides a powerful tool for such soft partitioning of the data set. Thus, people began to deal with clustering with fuzzy fashion and named them fuzzy cluster analysis. Since fuzzy clustering obtains the degree of uncertainty of samples belonging to each class and expresses the intermediate property of their memberships, it can more objectively reflect the real world. Thereby, it has become the main content of studies on cluster analysis. I n the last two

decades a great number of variations of fuzzy based clustering methods had evolved starting with the Fuzzy C-mean [2] in 1984.

Fuzzy clustering has been shown to be advantageous over crisp clustering in that the total commitment of a vector to a given class is not required in each iteration. Fuzzy methods have shown spectacular ability to detect not only hyper volume clusters, but also clusters which are actually thin shells that is curves and surfaces [x3]. We can see a number of examples in the literature [11]-[12] dealing with shell like volume curves and surfaces.

When compared to their crisp counterparts, fuzzy methods are more successful in avoiding local minima of the cost function and can model situations where clusters actually overlap.
The variants of FCM are themselves emerged from the FCM, as researchers have worked to improve its performance. If some researchers have worked to decrease the time consumptions others have worked to improve its accuracy. This method has been applied to various data types especially in the area of image segmentation with slight variations.

In outline, the clustering process for chemical structures is as follows.
(1) Select a set of attributes on which to base the comparison of the structures. These may be structural features and/or physicochemical properties. (2) Characterize every structure in the dataset in terms of the attributes selected in step one. (3) Calculate a coefficient of similarity, dissimilarity, or distance between every pair of structures in the dataset, based on their attributes. (4) Use a clustering method to group together similar structures based on the coefficients calculated in step three. Some clustering methods may require the calculation of similarity values between the new objects formed and the existing objects. (5) Analyze the resultant clusters or classification hierarchy to determine which of the possible sets of clusters should be chosen [x7]. Based on the work [x7] the wards crisp clustering method is the industry standard. They used MACCS search keys (MDL), Unity (Tripos) and Daylight 2D descriptors, Unity 3D rigid and flexible descriptors and two Abbott in-house 3D descriptors based on potential pharmacophore points. Further, they have compared Ward's and group average hierarchical agglomerative, Guénoche hierarchical divisive and Jarvis-Patrick non-hierarchical clustering methods. The results suggested that 2D descriptors and hierarchical clustering methods are best used for separating biologically active molecules from inactives. In particular, the combination of
MACCS descriptors and Ward's clustering was optimal.

2. Clustering Approaches

In order to derive the objective function and other relevant mathematics for fuzzy c-means and the remaining of its variations, it is better to see the same for the hard (crisp) partitioning technique, so that we may be able to understand the difference between the two approaches. (If we look into these issues all of them appears to be objective functional minimization problems. If the constrains are relaxed we get the possibilistic partition scheme. So, the clustering algorithm is nothing but a minimization problem which may be constrained or unconstrained.)

2.1 Hard Partitioning

These kind of methods are based on classical set theory and defines the presence or absence of a data point in a partition subset on strict logic, that is the object either belong to a subset or not. So, such kind of methods divides a data set strictly into disjoint subsets.
Let us suppose that we have a data set Z and the objective is to partition (group or cluster) it into C clusters. If we suppose that C is known as a priori, then the hard partition of Z is a set

$$\{A_i \mid 1 \le i \le c\} \subset P(Z) \rightarrow 1(a).$$

$$\bigcup_{i=1}^{c} A_i = Z \rightarrow 1(b).$$

$$\bigcap_{i=1}^{c} A_i = \varnothing \quad i \neq j \rightarrow 1(c).$$

$$\varnothing \subset A_i \subset Z \quad 1 \le i \le c \rightarrow 1(d).$$

$$U = \left[\mu_{ik}\right]_{cxn} \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \rightarrow 1(e)$$

Equation 1(a)-(d) give us the different properties of the set Z. Each cluster has distinct elements and the collection of all these disjoints sets is the set Z.

U is a matrix of the membership values for each element of set Z in each subset A_i. But one thing should be kept in mind that the value of

membership variable μ_{ik} is always either 1 or 0, if an element Z_k is a member of a set A_i then its membership is 1 otherwise 0. Thus the sum of all the membership values of a data point (or element of set Z) is always 1 as is evident from equation 2(a) and 2(b). Here it can also be argued that the sum of membership values of all the data points in a cluster can not be greater than the total number of the data points in a data set, as it is assumed that after a partition of the data set there should be at least two partitions. Because without at least two partition sets, there will be no partition at all. It is obvious from equation 2(c).

$$\mu_{ik} \in \{0,1\} \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \quad \rightarrow \quad 2(a)$$

$$\sum_{i=1}^{c} \mu_{ik} = 1 \quad 1 \le k \le n \quad \rightarrow \quad 2(b)$$

$$0 < \sum_{k}^{n} \mu_{ik} < n \quad 1 \le i \le c \quad \rightarrow \quad 2(c)$$

The biggest draw back of a hard partitioning is the concept that it either includes a data point in a partition or strictly excludes it; there is no other chance for the data elements to be part of more than one partition at the same time. However, in natural clusters it is always the case that some of the data elements partially belong to one set and partially to one or more other sets. In order to overcome this limitation, the notion of fuzzy partitioning was introduced.

2.2 Fuzzy Partitioning

Generalization of the fuzzy based partitioning can easily be followed from hard (crisp) partitioning if we allow the membership variable to attain any value between 0 and 1. The data points can now be considered to having partial memberships of more than one cluster. Equations for Fuzzy partitioning are analogous to 2 as given by Raspuni [4]:

$$\mu_{ik} \in [0,1] \quad 1 \le i \le c \quad 1 \le k \le n \quad \rightarrow \quad 3(a)$$

$$\sum_{i=1}^{c} \mu_{ik} = 1 \quad , \quad 1 \le k \le n \quad \rightarrow \quad 3(b)$$

$$0 < \sum_{i=1}^{n} \mu_{ik} < n \quad , \quad 1 \le i \le c \quad \rightarrow \quad 3(c)$$

One can observe that eqn 3 constrains the sum of all the membership values of a data point within all clusters prototypes equal to one. So, although the fuzzy partition allows the data points to have partial memberships but can not have total membership less than one.

2.3 Possibilistic Partitioning

The popularity of fuzzy set theory in the fields such as control and rule based reasoning is due to its ability to represent ill defined classes and concepts in a natural way [x3]. But again the fuzzy partitioning approach constrains the membership of any data point that membership values of a data points must sum to 1. It is the possibilistic approach that waives off this constraint and allows the data points to independently acquire any membership value in any cluster.

Equation 3(b) is replaced with a less restrictive constrain $\forall_\kappa, \exists_i, \mu_{ik} > 0$. So, the conditions for a possibilistic partition matrix are:

$$\mu_{ik} \in [0,1] \quad 1 \le i \le c \quad 1 \le k \le n \quad \rightarrow \quad 4(a)$$

$$\exists_i, \quad \mu_{ik} > 0, \quad \forall_\kappa, \quad \rightarrow \quad 4(b)$$

$$0 < \sum_{i=1}^{n} \mu_{ik} < n \quad , \quad 1 \le i \le c \quad \rightarrow \quad 4(c)$$

3. Algorithms

In the literature a big number of fuzzy algorithms can be found which are based on the two partitioning concepts i.e fuzzy and possibilistic. In this section a number of algorithms like fuzzy c-mean, G-K, Gath and Geva, and some other newly optimized.

3.1. Fuzzy c-mean

The fuzzy c-mean is the basis of all its fuzzy based clustering variants. It is based on the minimization of an objective functional iteratively.

$$J(C,U,Z) = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m \| Z_j - C_i \|^2_A \quad \rightarrow \quad 5(a)$$

Here Z_j is the jth data element in the dataset, C_i the prototype of the ith cluster, μ_{ij} is membership value for the jth data element in ith cluster, and m is the fuzzification parameter whose typical value can be from 1.1 to 2.2. The following steps are taken iteratively to reach the absolute minimum:

Step1 Initialize *m*, the partition matrix *U*, and the number of clusters *C*.

Repeat the following steps

Step2 Compute the cluster prototypes

$$V_i^{(t)} = \frac{\sum_{k=1}^{n} (\mu_{ik}{}^{(t-1)})^m Z_k}{\sum_{k=1}^{n} (\mu_{ik}{}^{(t-1)})^m} \quad 1 \le i \le c \quad \rightarrow \quad 6(a)$$

Step3 Compute the distances

$$D^2_{ikA} = (Z_k - V_i^{(l)})^T A (Z_k - V_i^{(l)}) \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \quad \rightarrow \quad 6(b)$$

Step4 Update the partition matrix
If $D_{ikA} > 0$

$$\mu^{(l)}_{ik} = \frac{1}{\sum_{j=1}^{c} (D_{ikA} / D_{jkA})^{2/(m-1)}} \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \quad \rightarrow \quad 6(c)$$

else $\mu_{ik} = 0$

Step5 if $\left\| U^{(l)} - U^{(l-1)} \right\| < \varepsilon$ **Stop**

Else go to Step 2

3.2 Guftafson & Kessel Algo

In order to overcome the circular or same size and shape clusters of the FCM, Gustafson and Kessel [x8] introduced the notion of covariance matices (F_i) in distance calculation. The matrix A in equation 6(b) is replaced with the covariance matrix for each cluster i.

$$F_i = \frac{\sum_{k=1}^{n} (\mu_{ik}^{(l-1)})^m (Z_k - V_i^{(l)})^T (Z_k - V_i^{(l)})}{\sum_{k=1}^{n} (\mu_{ik}^{(l-1)})^m} \quad 1 \le i \le c \quad \rightarrow \quad 6(d)$$

and the distance:

$$D^2_{ikA} = (Z_k - V_i^{(l)})^T [\delta \det(F_i)^{1/n} F_i^{-1}] (Z_k - V_i^{(l)}) \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \quad \rightarrow \quad 6(e)$$

3.3 Possibilistic FCM

According to this approach the membership values are calculated as follows:

$$\mu_{ik} = \frac{1}{1 + (D^2_{ik} / \eta_i)^{1/m-1}} \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \quad \rightarrow \quad 6(f)$$

and

$$\eta_i = K \frac{\sum_{k=1}^{n} \mu_{ik}^m D^2_{ik}}{\sum_{k=1}^{n} (\mu_{ik})^m} \quad \rightarrow \quad 6(g)$$

4. Discussion

The fuzzy based clustering methods had shown tremendous achievements in areas of image processing and pattern recognition. The fuzzy c-mean is a good choice for circular or spherical clusters. But if the orientation of natural clusters are not spherical then the algorithms leads to almost wrong clusters. Another drawback of the algorithm is that it imposes equal size clusters on the data set which is again a deviation from the natural clusters. The GK algorithm partly improves the scenario, by using a different norm matrix ie covariance matrix. But due to a mathematical constraint in the algorithm the determinant of the norm matrix have to be equal to some constant, which imposes the constant volume clusters but with the advantage of any cluster can now have any kind of shapes, may be spherical, ellipsoidal, linear or shells. In [x1] they have shown with example how accurately GK clusters the almost very close rectangular clusters. In [x2] it has been shown that the same variable shape clusters can be achieved with FCM if instead of Euclidean distance, Mahalanobis distance is used with some modification. According to [x3] PCM has the ability to filter out noise and outliers from the main natural clusters but at some time it goes into the problem of coincident clusters. In [x4] PCM has been used to cluster a multichannel satellite images but the performance shows very poor results coincident clusters. The problem in PCM is that the cluster centers are also attracted to one another as the elements of the clusters are attracted towards its centre. This problem has been reduced by incorporating cluster repulsion, a technique in which the cluster centers are stretched away from one another [x5]. The performance of any fuzzy based clustering method is the best when the number of clusters is known a priori. But most of the time it is not the case and so researchers has devised a number of methods known as cluster validation indices to evaluate the clusters formed [33, 34, 36, 37 and 38].

The clustering of chemical datasets is not very new but fuzzy based methods have not yet been employed to their fullest performance. In [x6] Rodgers et al have extensively used the fuzzy k-means for clustering files of chemical structures and has compared its performance with the established crisp methods like simple c-mean and wards. There results show a little advantage over the crisp methods but at very small values of the fuzzification parameter m. That can not be interpreted as a good mark for the fuzzy c-mean in clustering chemical structures.

5. Conclusion

The applications of fuzzy based methods in all fields of engineering and sciences have shown far reaching results and their applications in chemo informatics is also optimistic. In [x6] an average performance enhancement of around 6% is achieved. Thus there is a need to adapt other fuzzy clustering variants to clustering of chemical datasets.

6. References

[1] Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. “Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial

Libraries", *J. Med. Chem*., **1994**, *37*, 1233-1251.

[2] Hecht, P. "High-throughput screening: beating the odds with informatics-driven chemistry", *Curr. Drug Discov*., January 2002, 21-24.

[3] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: Fuzzy c-means algorithm", Computers and Geoscience, 1984.

[4] R. Krishnapuram, J.M. Keller, "A possibilistic approach to clustering**",** IEEE Transactions on Fuzzy Systems, Volume: 1, Issue: 2, May 1993.

[5] R. N. Dave, "Fuzzy shell-clustering and applications to circle detection in digital images," Int. J. General Systems, vol. 16, pp. 343-355, 1990

[6] R. Krishnapurum, O. Nasraoui, and H. Frigui, "The Fuzzy C-shells algorithm: A new approach", IEEE Transaction on Neural Networks, vol. 3, No. 5, pp. 663-671, Sep. 1992.

[7] R. D. Brown, and Y. C. Martin, "Use of structure- Activity data to compare structure based clustering methods and descriptors for use in compound selection", J. chem. Info. and comp. science, 36, 572-584. 1996.

[8] E. Ruspini, "Numerical methods for fuzzy clustering", *Inf. Sci. 2*, 319–350, 1970.

[9] Gustafson, D.E. and W.C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix" *Proc. IEEE CDC*, San Diego, CA, USA, pp. 761–766. 1978.

[10] F. Hopner, F. Klawonn, R. Kruse and T. Runkler. "Fuzzy Cluster Analysis" pp. 43-45, John Wiley & Sons, 1999.

[11] R. Krishanpurum and J. Kim, "A note on Gustafson - Kessel and Adaptive Fuzzy Clustering Algorithm", IEEE Transaction on Fuzzy Systems, vol.7, no. 4, August 1999.

[12] M. Barni, V. chappelini and A. Mecocci, "Comments on apossibilistic approach to clustering" IEEE Transaction on Fuzzy Systems, Vol. 4 No.3, August 1996.

[13] H. Timm, C. Borgelt, C. Doring and R. Kruse, "An extension to possibilistic fuzzy cluster analysis", Fuzzy Sets and Systems, vol. 147, pp. 3-16, 2004.

[14] D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 1, pp. 224-227, 1979.

[15] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," J. Cybernetics, vol. 3, pp. 32-57, 1973.

[16] R.B. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," Comm. in Statistics, vol. 3, pp. 1-27, 1974.

[17] U. Maulik, S. Bandopadhyay, "performance Evaluation of some clustering Algorithms and Validity Indices", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 24, No. 12, December, 2002.

[18] J.C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 (1974) 57–71.

[19] J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybernet. 3 (1974) 58–72.

[20] J. D. Holliday, SW. L. Rodgers, P. Willet, et all, "Clustering Files of chemical Structures Using the Fuzzy k-means Clustering Method", J. chem. Info. and comp. science, 44, 894-902. 2004.