A Soft Hierarchical Algorithm for the Clustering of Multiple Bioactive Chemical Compounds

Jehan Zeb Shah* and Naomie bt Salim

Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor Darul Ta'zim, Malaysia
zebjehan@gmail.com, naomie@fsksm.utm.my

Abstract. Most of the clustering methods used in the clustering of chemical structures such as Ward's, Group Average, K- means and Jarvis-Patrick, are known as hard or crisp as they partition a dataset into strictly disjoint subsets; and thus are not suitable for the clustering of chemical structures exhibiting more than one activity. Although, fuzzy clustering algorithms such as fuzzy c-means provides an inherent mechanism for the clustering of overlapping structures (objects) but this potential of the fuzzy methods which comes from its fuzzy membership functions have not been utilized effectively. In this work a fuzzy hierarchical algorithm is developed which provides a mechanism not only to benefit from the fuzzy clustering process but also to get advantage of the multiple membership function of the fuzzy clustering. The algorithm divides each and every cluster, if its size is larger than a pre-determined threshold, into two sub clusters based on the membership values of each structure. A structure is assigned to one or both the clusters if its membership value is very high or very similar respectively. The performance of the algorithm is evaluated on two bench mark datasets and a large dataset of compound structures derived from MDL's MDDR database. The results of the algorithm show significant improvement in comparison to a similar implementation of the hard c-means algorithm.

Keywords: cluster analysis, chemoinformatics, fuzzy c-means, bioinformatics, chemical information systems.

1 Introduction

The clustering of drug like compound structures is important in many phases of drug discovery and design like the virtual screening, prediction and modeling of structure properties, virtual library generation and enumeration etc. Drug discovery is a complex and costly process, with the main issues being the time and costs of finding, making and testing new chemical entities (NCE) that can prove to be drug candidates. The average cost of creating a NCE in a major pharmaceutical company was estimated at around $7,500/compound [1]. For every 10,000 drug candidate NCE synthesized, probably only one will be a commercial success and there may be 10-12 years after it is first synthesized before it reaches the market [2].

* Corresponding author.

S. Hochreiter and R. Wagner (Eds.): BIRD 2007, LNBI 4414, pp. 140 – 153, 2007.

Currently, many solution- and solid- phase combinatorial chemistry (CC) strategies are well developed [3]. Millions of new compounds can be created by these CC based technologies but these procedures have failed to yield many drug candidates. Enhancing the chemical diversity of compound libraries would enhance the drug discovery. A diverse set of compounds can increase the chances of discovering various drug leads and optimization of these leads can lead to better drugs. In order to obtain a library of high chemical diversity, a number of structural processing technologies such as diversified compound selections, classification and clustering algorithms have been developed. However, the need for more robust and reliable methods is still seriously felt [4].

The term cluster analysis was first used by Tryon in 1939 that encompasses a number of methods and algorithms for grouping objects of similar kinds into respective categories [5]. The main objective of clustering 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. There is no pre-notion about the groups present in the data set.

Willett [6] has found that, among the hierarchical methods, the best result was produced by Ward's, Group Average and Complete Linkage hierarchical methods and Jarvis-Patrick was found to be the best method among the non-hierarchical methods tested. They have evaluated almost 30 hierarchical and non hierarchical methods on 10 datasets each containing a group of compounds exhibiting a particular property or biological activity such as anesthetic activity, inhibition activity, molar refractivity, where 2D fingerprints been used as compound descriptors. In another study [7], Barnard and Downs have further investigated Jarvis-Patrick method in more detail using a small dataset of 750 diverse set of compounds from the ECDIN database using 29 physiochemical and toxicological information. Though satisfactory correlations have been obtained yet to obtain the best correlations for different properties and activities different parameter setting was necessary.

In [8], Downs and Willett have analyzed the performance of Ward's, Group Average, Minimum Diameter and Jarvis Patrick methods on two datasets: a small subset of 500 molecules and another one of 6000 molecules from Chemical Abstract Service [9] database. They have incorporated the same 29 physiochemical properties. The performance of Jarvis Patrick's method was very poor. The Minimum diameter method was found to be the most expensive, and the performance of the Ward's method was the best.

Another work on the clustering of chemical dataset was reported by Brown and Martin [9] where Ward's, Jarvis-Patrick's (fixed and variable length nearest neighbor lists), Group Average and Minimum Diameter (fixed and variable diameter) methods have been evaluated on four datasets, each with single activity containing active as well as inactive compounds, summing to a total of 21000 compounds. They have employed a number of descriptors like MACCS 2D structural keys, Unity 2D keys, Unity 2D fingerprints, Daylight 2D fingerprints and Unity 3D pharmacophore screens. The performance of wards was found to be the best across all the descriptors and datasets, whilst the Group average and minimum diameter methods were slightly inferior. The performance of Jarvis Patrick method was very poor for fixed as well as for variable length nearest neighbor lists.

Recently, fuzzy clustering methods have been applied for clustering of chemical datasets. In [10] Rodgers et al have evaluated the performance of fuzzy c-means algorithm in comparison with hard c-mean and Ward's methods using a medium size compound dataset from Starlist database for which LogP values were available. Their results show that fuzzy c-means is better than Ward's and c-means. They have used simulated property prediction method [11] as performance measure, where the property of the cluster is determined by the average property of all the molecules contained in the cluster. This average property of the cluster is called the simulated property of each of the structure in the cluster. The simulated property of each molecule is correlated with the actual property of the compound to find the performance. In [12], we have used fuzzy Gustafson-Kessel, fuzzy c-means, Ward's and Group Average methods to cluster a small size dataset from MDL's MDDR database containing about 15 biologically active groups. Instead of using simulated property prediction method, the active cluster subset method where the proportion of active compounds in active clusters is used as performance measure, was employed. Our results show that the performance of Gustafson-Kessel algorithm is the best for optimal number of clusters. The Ward's, fuzzy c- means and Group Average methods are almost the same for optimal number of clusters.

Bocker et al [13] have revised the hierarchical k-means algorithm and developed an interface to display the resultant hierarchy of compound structures in the form of a very useful colorful dendrogram. The same system has also been used for the display of results for an improved median hierarchical algorithm [14].

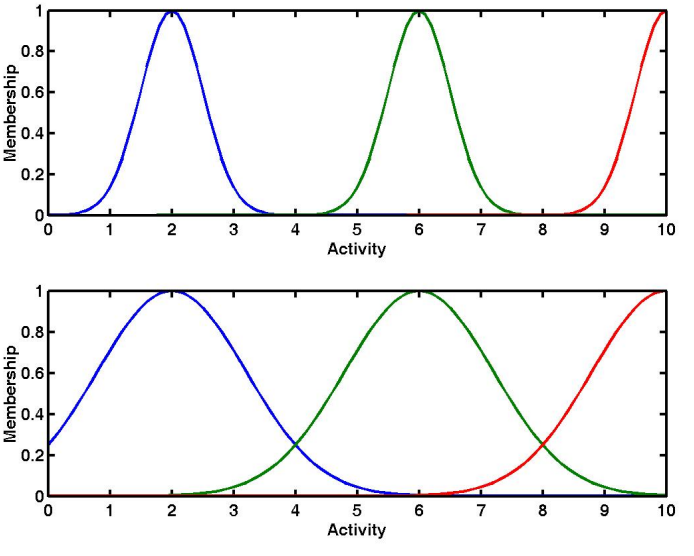

Fig. 1. The upper graph describes the non-overlapping compound structures whereas where as the lower graph describes the overlapping clusters. The vertical axis plots the strength of activity of the compound structure and horizontal axis plots the number of activities.

The main problem of the current clustering methods used in the field of chemoinformatics is their non overlapping nature, where these methods consider the datasets as having very distinct and clearly separable boundaries among the various clusters. It is contradictory to the real problems at hand, where the boundaries are very vague and so it is not always as simple to delineate the clusters as these traditional methods tackle them. In Figure 1, two types of datasets are depicted a) where the boundaries of the clusters are well defined and can easily be separated and b) where the boundaries of the clusters are vague and so difficult to delineate. The second case is a challenge for the current methods where their performance is expected to be not as good as when each data element is not limited to belong to only one cluster.

In case of chemical compounds that are biologically active, it is often the scenario that they exhibit more than one activity simultaneously and grouping such compounds under one cluster is not justified. For example, the MDDR database [15], which currently contains around 16 million compounds, the number of compounds active against multiple targets is considerably large.

Thus the previous researches show the effectiveness of hierarchical methods on one hand and that of the fuzzy methods on the other hand. The fuzzy clustering methods have the promise to care for the overlapping nature of chemical structures. In this work an improved hierarchical fuzzy algorithm is employed for the clustering of chemical structures that exhibit multiple activities. It has also been shown on a dataset sufficiently quantified in terms of the activities that the method result in better clustering when higher overlapping is allowed in the clustering process. The results have also been compared with a similar implementation of the k-means method.

In the next section the dataset and the corresponding structural descriptors used in this work are described. Section 3 discusses the hierarchical implementation of the fuzzy c – means in detail. In section 4 the results are discussed and section 5 concludes the work.

2 Dataset Preparation and Descriptors Generation

In this work three datasets have been utilized to evaluate the performance of the proposed algorithm: two benchmark datasets known as the Fisher's Iris dataset [16] and Golub's Lekuemia datasets [17] and one drug dataset composed of bioactive molecules exhibiting overlapping as well as non-overlapping activities collected from the MDDR database. The MDDR database contain 132000 biologically relevant compounds taken from patent literature, scientific journals, and meeting reports [15]. Each entry of the database contain a 2D molecular structure field, an activity class and an activity class index fields besides many other fields like biological investigation phase, chemical abstract service (CAS) [18] compound identity , and patent information fields. The activity index is a five digit code used to organize the compounds' structures based on biological activity; for example the left most one or two digits describe a major activity group and the next three digits describes sub activities inside the bigger activity. For example, the activity index 31000 shows a large activity of antihypertensive agents and the activity indexes 31250, 31251 show Adrenergic (beta) Blocker, Adrenergic (beta1) Blocker respectively. The dataset used

here comprised of 12 major activities where each group can further be divided into a few sub categories. Initially 55000 compounds have been extracted from the database using a number of filtering strategies (as described in the following equations 1 and 2). The number of compounds in dataset1 (DS1) was 29843 and dataset2 (DS2) 6626. The dataset DS1 contain exactly non-overlapping structures where each compound in the dataset can exhibit only and only one activity among the list of activities selected for this work. The dataset DS2 contain bioactive (compounds exhibiting only two activities) such that each compound exhibit two activities only. Let ***A*** be the set of activities selected and ***l*** be the number of activities in this set, then the DS1 is a superset of sets $\boldsymbol{D_i}$, a set of compounds exhibiting activity $i \in A$. If $z \in DS1$ is an arbitrary compound, then

$$DS1 = \left\{ z \in D_i \mid \quad i \in A \wedge i \notin A - \{i\}, \quad i = 1,2,...,l \right\} \tag{1}$$

Similarly, DS2 is a superset of sets $\boldsymbol{D_{ij}}$, where compound $z \in DS2$ exhibit two activities $i, j \in A$ and so,

$$DS2 = \left\{ z \in D_{ij} \mid \quad i, j \in A \wedge i, j \notin A - \{i, j\} \quad \begin{matrix} i = 1,2,...,l-1 \\ j = i+1,...,l \end{matrix} \right\} \tag{2}$$

By combining these two datasets, another dataset DS3 has been organized in the same way as depicted in fig 1. It contains single activity compounds from DS1 and in between any two activity groups there are bi-activity molecules from DS2 which will belong to both the groups on its right and left.

The descriptors generation or features extraction is an important step in computational clustering of molecular structures and other problems such as classification and quantitative property/activity relationship modeling. A number of modeling tools are available that can be used to generate structural descriptors. In this work, we use the Dragon software [19] to generate around 99 topological indices for the molecules. Topological indices are a set of features that characterize the arrangement and composition of the vertices, edges and their interconnections in a molecular bonding topology. These indices are calculated from the matrix information of the molecular structure using some mathematical formula. These are real numbers and possess high discriminative power and so are able to distinguish slight variations in molecular structure. This software can generate more than 1600 descriptors which include connectivity indices, topological indices, RDF (radial distribution function) descriptors, 3D-MORSE descriptors and many more.

Scaling of the variables generated is very important in almost all computational analysis problems. If magnitude of one variable is of larger scale and the other one is of smaller scale then the larger scale variable will dominate all the calculations and effect of the smaller magnitude variables will be marginalized. In this work all the variables used were normalized such that the maximum value for any variable is 1 and the minimum is 0.

Table 1. Selected Topological Indices

TI	Description
Gnar	Narumi geometric topological index
Hnar	Narumi harmonic topological index
Xt	Total structure connectivity index
MSD	Mean square distance index (Balaban)
STN	Spanning tree number (log)
PW2	path/walk 2 – Randic shape index
PW3	path/walk 3 – Randic shape index
PW4	path/walk 4 – Randic shape index
PW5	path/walk 5 – Randic shape index
PJI2	2D Petitjean shape index
CSI	Eccentric connectivity index
Lop	Lopping centric index
ICR	Radial centric information index

In order to reduce the descriptor space and to find the more informative and mutually exclusive descriptors a feature selection method principal component analysis (PCA) [20] was used. PCA was carried out using the MVSP 3.13 [21]. It has been found that 13 components can represent more than 98% of the variance in the dataset. The input to our clustering system is thus a 13 X 28003 data matrix. The 13 selected topological indices are shown in Table 1.

3 Methods

Fuzzy clustering is the intrinsic solution to the problem of overlapping data, where the data elements can be member of more than one cluster. The traditional clustering methods do not allow this shared membership by restricting the data elements to belong to only one of the many clusters exclusively. There can be almost three types of partitioning concepts, the traditional hard or crisp one where a compound can belong to only one cluster and so the membership degree of the compound is said to be 1 in any one cluster and zero in the rest of the clusters. Another approach is provided by the fuzzy logic where the membership degree of a compound can be [0, 1] and so the compound can belong with varying degree to more than one cluster [22-23]. In both of these partitioning scenarios, the membership μ_{ik} follow a few conditions such as the sum of the membership values over a range of clusters c is always equal to one:

$$\sum_{i=1}^{c} \mu_{ik} = 1 \quad 1 \le k \le n \tag{3}$$

$$0 < \sum_{k=1}^{n} \mu_{ik} < n \quad 1 \le i \le c \tag{4}$$

Where n is the number of compounds in the dataset, c is the number of clusters and i and k are the indexes for the clusters and data elements respectively.

In the third case, called the possibilistic partitioning [24], this constrain is also relaxed and the sum of the membership degrees is not required to be equal to one, however, clustering algorithms based on this theory are out of the scope of this work.

3.1 Fuzzy Clustering Algorithm

In the literature there are a large number fuzzy based clustering algorithms [22-26] that are variations of the most fundamental and widely used fuzzy c-means, a fuzzy counter part of the ordinary c-means (or k-means) algorithm, first characterized by Dunn [27] and then formalized by Bezdek [28]. The algorithm is based on the iterative minimization of an objective functional and so independent of the initialization conditions and sequence of input presentation. The objective functional is given as:

$$J(C,U,Z) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m \left\|Z_k - V_i\right\|^2 A \tag{5}$$

where Z_k is the k^{th} feature vector representing a molecule of the dataset Z containing a total of n molecules, V_i is the prototype (center or codebook) of i^{th} cluster of the total number of clusters c, $\|.\|^2$ is the square of distance between each molecule and each cluster center, and μ_{ik} is the membership value of molecule Z_k to be part of prototype V_i. A represents a positive definite norm inducing matrix dependent on the type of distance (in case of Euclidean distance it is a unity matrix). The stepwise description of the algorithm is given below:

Step1. Initialize the fuzzification index m, the partition matrix U, the number of clusters c and tolerance ε.

Repeat the following steps

Step2. Compute the cluster prototypes

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

where the subscripts l and l-1 represent the current and previous iterations respectively.

Step3. Compute the distances between compound Z_k and cluster center V_i

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

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

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

else

$$\mu_{ik} = 0 \tag{6d}$$

Step5. if $\left\| U^{(l)} - U^{(l-1)} \right\| < \varepsilon$ **Stop**
else go to Step 2

The fuzzification parameter m is a measure of the fuzzfication that can have any value from 1.1 to ∞. As the value of m is increased the memberships of molecular structures to the clusters become fuzzier. For a value of $m = 1$, the algorithm will simply becomes the crisp or hard c-means, but it should be avoided as it will result in a divide by zero error in equation 6(c). Many researches suggest a value of 2.0 for m, as the first fuzzy c-means algorithm suggested by Dunn also used the same value [29]. The stopping condition $\varepsilon = 0.001$ is usually enough for convergence, but we have kept it at 0.0001, just to be on the safe side.

3.2 Fuzzy Hierarchical Clustering

The algorithm is a recursive procedure of fuzzy clustering, where each cluster formed is further re-clustered. The number of child clusters in each recursive call can be 2, 3 or any other number greater than 1. However, here in every recursive call the value of c is kept at 2 to obtain binary tree like order on the structures, a fashion more suitable and historical to the chemical structures based on their biological activities. The inputs are a n *X* m data matrix Z composed of n (number of structures in each recursive call) rows of feature vectors $Z_k \in \Re^m$ and m columns of features. The output of each recursive call is a c *X* n membership matrix U. The two child clusters are formed using the membership matrix U, where a structure Z_k can be a member of either one of the two clusters if the membership of one is greater than the other to some extent, or can be part of both the clusters if their membership degrees do not show much difference. Once a cluster is partitioned into its child clusters, the membership matrix is discarded but the algorithm keeps the necessary global information in the constituent clusters by adding the structures which are closely related to both the clusters. Thus in each recursive call a new membership matrix is generated and optimized based on the local information of the cluster.

This recursive process of clustering continues until every cluster is a singleton (a cluster containing only one structure) or when an optimal partition is obtained. For this purpose the partition validity measure suggested by Bocker et al [13] is adopted. The clustering process is repeated for a number of threshold and at the end of each repetition, the number of singletons, the number of non singleton clusters, and a

distance measure D_{max} (Equation 7) are calculated and plotted against the thresholds to find the optimal threshold.

$$D_{\max} = \sum_{i} \max[d(z_k, c_i)], \quad 1 \le k \le n \tag{7}$$

where d is the Euclidean distance, between the structure $z_k \in c_i$ and the prototype of the cluster c_i , and n is the number of structures in each cluster. The value of D_{max} represents the maximum deviation of the clusters from their prototypes.

Once the optimal threshold is obtained from the graph by visual inspection (one shown in Fig 2(a)-(b)), the clustering process is repeated for the last time with the best threshold selected. The main steps of the algorithms are ordered below:

Run1: For finding the optimal threshold

(i) A threshold is selected from a range of thresholds
(ii) The value c is initialized which is 2 for binary trees, the membership matrix U is initialized
(iii) The dataset is clustered using the fuzzy –c-means algorithm
(iv) Each of the cluster is checked if the size is larger than the Threshold selected, then go to step (ii) for sub-clustering the resultant cluster
(v) Plot the number of clusters, singletons and the metric D_{max} against the range of threshold

Run2: (i) Select the optimal threshold through visual inspection of the graph
(ii) Repeat the algorithm for the last time using the optimal threshold.

In clustering a good method is supposed to combine highly similar activity structures together, so large number of singletons is not considered a good gesture. Thus, an appropriate point for a good clustering will be a threshold for which the number of singletons is a minimum.

4 Results and Discussion

Three datasets have been used to evaluate the performance of the clustering process. These include two small size benchmark datasets a leukemia cancer dataset and fisher iris dataset, and a real dataset of chemical structures described earlier in detail. Leukemia dataset is a collection of 72 genes expressions belonging to two types of cancer, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).Almost 62 of the specimens for this genes expression data were obtained from bone marrow samples of the acute leukemia patients while the rest had been collected from the peripheral blood samples. The fisher's Iris dataset consists of 150 random samples of flowers belonging to the Iris species setosa, versicolor, and virginica. For each of the specie, the dataset contain 50 samples and each sample consists of four variables, sepal length, sepal width, petal length and petal width.

The Iris dataset poses much difficulty to be partitioned into three classes as two of the classes are highly overlapped [30, 31]. However, our method can partition the dataset into three clusters with high accuracy, when a good threshold is selected.

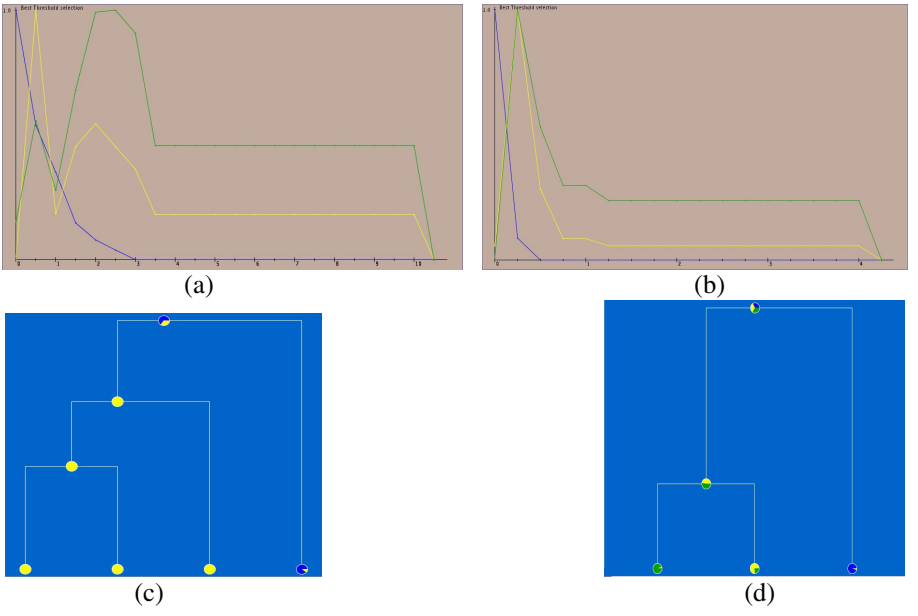

Fig. 2. Clustering results of (a) lekuemia threshold, (b) iris threshold (c) leukemia clustering tree and (d) iris clustering tree

The threshold selection plot is given in figure 2(b) and corresponding dendrogram for the threshold value 0.5, is given in figure 2(d). It can easily be observed that one (shown in yellow color) of the classes could not been separated with much accuracy.

Since these two real datasets are almost non-overlapping but when the number of clusters is decreased, the accuracy (performance) of the clustering degrades. These two results are for the threshold level 0.5 (iris) and 3.0 (leukemia). As the threshold is decreased, the clustering accuracy increases but results in more number of clusters and as the threshold is increased lesser number of clusters and more heterogeneous clusters are obtained.

After confirming the results with the help of benchmark datasets, the methods was applied to the real molecular dataset DS3 developed in section 2. This dataset contain around 12 biologically active and overlapping clusters and the objective of the work is to evaluate the clustering performance of the developed hierarchical fuzzy c-mean (HFCM) algorithm. For evaluation, we use the active cluster subset method [9]. A threshold range of 0.01-0.1 with an increment step of 0.01 was used in this work. For each threshold a number of clusters were obtained. Some of the clusters obtained may be having only actives or inactives structures but many of them will have both. The clusters having at least one active structure are combined to make one super cluster called the active cluster subset.

This subset of the dataset used should not contain any of the singletons, the singletons do not give any clue about the performance of the clustering method, and the clustering method should combine active structures with actives and inactives

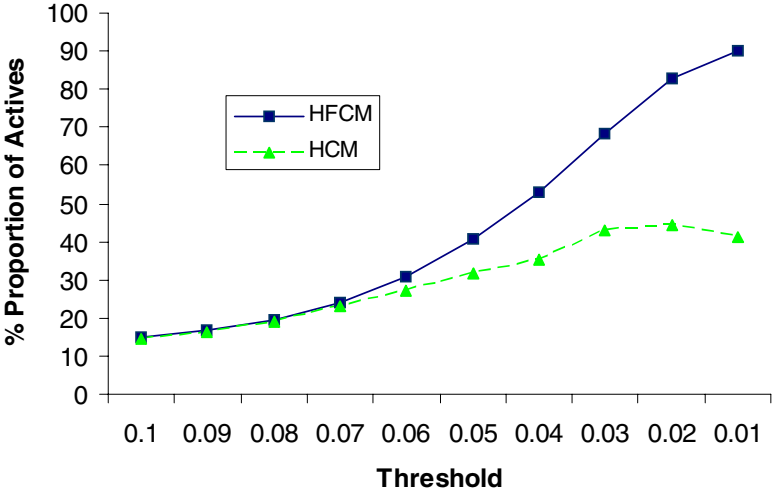

Fig. 3. Performance of the Hierarchical Fuzzy c-means (HFCM) and hierarchical c-means (HCM)

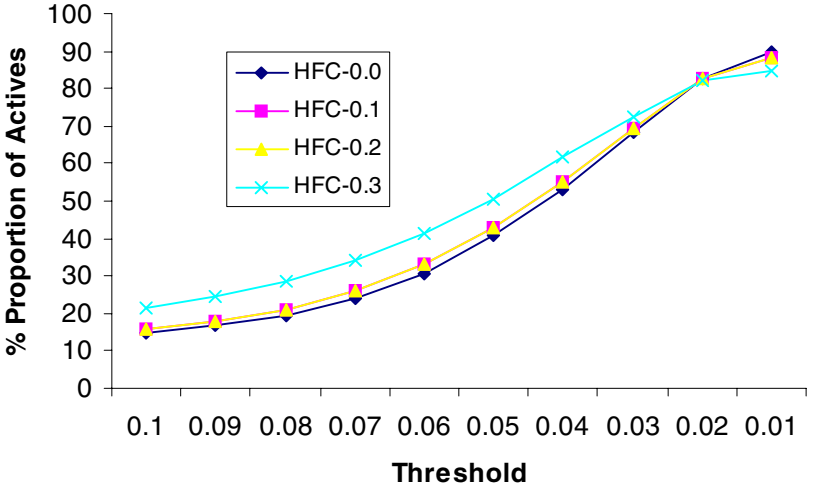

Fig. 4. Performance of the HFCM for various Membership Thresholds. The HFC-0.0 stands for the HFCM with Membership Threshold of 0.0, and so on.

structures with inactives. Thus the singletons are avoided to be included in the active cluster subset. The proportion of actives to inactive structures in the active cluster subset is determined. For each activity group of the dataset, the structures belonging to that activity group were taken as active and the rest of the groups were taken as inactive. The process is repeated for all the 12 bioactivity groups of the dataset for all the clusters obtained for each threshold level and an average proportion was determined.

The fuzzification index of the fuzzy c-means determines the spread in the dataset [24] whose value can range from 1.1 to any finite number, a smaller value means that the data and natural clusters are spread over wide area (volume). As the value of fuzzification index is increased the data and the clusters within becomes more and more compact. The best value of the fuzzification index for which the best clustering can be obtained depends on the dataset used. So, first a fuzzy c-mean method was used to determine the best value for fuzzification index, and was found to be 1.4. The performance of the hierarchical fuzzy c-means is shown in figure 3, for various values of the threshold level, in comparison with the hierarchical hard c-means clustering. Further, to investigate the effect of overlap we repeated a number of experiments. The parent cluster was partitioned into two child clusters based on the membership degree of each compound structure as follows:

If $U_{ik} - Threshold > 0$

The compound k is assigned to cluster i, $0 <= i <= 1$

else

the compound is assigned to both of the two clusters

Since, $\mu_{ik} \in [0,1]$, so, the value of *Threshold* can be between 0 and 0.5. We have tested for a number of values of *Threshold* and the results are shown in figure 4. As the *Threshold* is increased the compounds are allowed to show more overlap and so we get compounds that go to both of the two child clusters. This permission of higher overlap results in small size and homogeneous clusters, which increases the percentage of active structures in active cluster subset.

5 Conclusions

In this work, an improved hierarchical fuzzy algorithm has been employed for the clustering of chemical structures. The results of the algorithm are very convincing in clustering the multiple activity compounds. A special real dataset have been developed for this purpose where the overlap of activities have been restricted to only two which complements the analysis process for binary tree like clustering. It has been shown that the algorithm have an edge over a similar implementation of the k – means algorithm. Moreover, when higher overlap of activities is allowed, which is incorporated by fuzzy membership as threshold, the results are improved.

References

1. P. Hecht, "High-throughput screening: beating the odds with informatics-driven chemistry," Current Drug Discovery, 2002, pp. 21-24.
2. W.A. Warr, High-Throughput Chemistry: Handbook of Chemoinformatics, Vol. 4, Wiley-VCH, Germany, 2003.

3. D.G. Hall, S. Manku, and F. Wang, Solution- and Solid-Phase Strategies for the Design, Synthesis, and Screening of Libraries Based on Natural Product Templates: A Comprehensive Survey, Journal of combinatorial Chemistry 3 (2001), 125-150.
4. C.N. Parker, C.E. Shamu, B. Kraybill, C.P. Austin, and J. Bajorath, Measure, mine, model, and manipulate: the future for HTS and chemoinformatics? , Drug Discovery Today 11:(19-20) (2006), 863-865
5. R.C. Tryon, Cluster Analysis, MI: Edwards Brothers, 1939.
6. P. Willett, Similarity And Clustering In Chemical Information Systems, Research Studies Press, Letchworth, 1987.
7. G.M. Downs and J.M. Barnard, Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures, Journal of chemical information and computer science 32:(6) (1992).
8. G.M. Downs, P. Willett, and W. Fisanick, Similarity searching and clustering of chemical structure databases using molecular property data, Journal of Chemical Information and Computer Science 34: (1994), 1094-1102.
9. 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, Journal of chemical Information and computer science 36 (1996), 572-584.
10. J.D. Holliday, S.L. Rodgers, and P. Willet, Clustering Files of chemical Structures Using the Fuzzy k-means Clustering Method, Journal of chemical Information and computer science 44 (2004), 894-902.
11. G.W. Adamson and J.A. Bush, A comparison of some similarity and dissimilarity measures in the classification of chemical structures, Journal of chemical Information and computer science 15 (1975), 55-58.
12. J.Z. Shah and N. Salim, FCM and G-K clustering of chemical dataset using topological indices, Proc of the First International Symposium on Bio-Inspired Computing, Johor Bahru, Malaysia, 2005.
13. A. Bocker, S. Derksen, E. Schmidt, A. Teckentrup, and G. Schneider, A Hierarchical Clustering Approach for Large Compound Libraries, Journal of chemical Information and modeling 45:(4) (2005), 807-815.
14. A. Bocker, G. Schneider, and A. Teckentrup, NIPALSTREE: A New Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening, Journal of chemical Information and computer science (2006).
15. "MDL's Drug Data Report," Elsevier MDL. http://www.mdli.com/products/knowledge/drug_data_report/index.jsp
16. R.A. Fisher, The use of multiple measurements in axonomic problems, Annual Eugenics 7 (1936), 179-188.
17. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science Magazine 285 (1999), 531 - 537.
18. "Chemical Abstract Service," website: http://www.cas.org/.
19. "Dragon," melano chemoinformatics. http://www.talete.mi.it
20. I. Jolife, Principal component analysis, Springer-Verlag, New York, 1986.
21. "MVSP 3.13, Kovach computing services: ." http://www.kovcomp.com/
22. J.C. Bezdek and R.J. Hathaway, Numerical convergence and interpretation of the fuzzy c-shells clustering algorithm, IEEE Transaction on Neural Networks 3, (1992), 787 – 793.
23. R.N. Dave, Fuzzy shell-clustering and applications to circle detection in digital images, International Journal of General Systems 16 (1990), 343-355.

24. F. Hopner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis, John Wiley & Sons, 1999.
25. R. Krishnapurum, O. Nasraoui, and H. Frigui, The Fuzzy C-shells algorithm: A new approach, IEEE Transaction on Neural Networks 3:(5) (1992), 663-671.
26. Y.H. Man and I. Gath, Detection and separation of ring-shaped clusters using fuzzy clustering, IEEE Transaction on pattern analysis and machine intelligence 16:(8) (1994), 855 – 861.
27. J.C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3 (1973), 32-57.
28. J.C. Bezdek, R. Ehrlich, and W. Full, FCM: Fuzzy c-means algorithm, Computers and Geoscience (1984).
29. H. Choe and J.B. Jordan, On the optimal choice of parameters in a fuzzy c-means algorithm, Proc of the IEEE Conference on Fuzzy Systems1992, pp. 349 - 354
30. I. Gath and A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Transaction on pattern analysis and machine intelligence 11:(7) (1989), 773-781.
31. A.B. Geva, Hierarchical unsupervised fuzzy clustering, IEEE Transaction on Fuzzy Systems 7:(6) (1999), 723-733.