Proceedings of the 1st International Conference on Natural Resources Engineering & Technology 2006
24-25th July 2006; Putrajaya, Malaysia, 492-499

Optimized Subtractive Clustering for Cluster-Based Compound Selection

Kuik Sok Ping, Naomie bt Salim*

Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.

Abstract

Compound selection algorithm has become a need to pharmaceutical industry due to the increasing number of chemical compounds to be screened. One of the widely used methods in compound selection is cluster-based selection where the compound datasets are grouped into clusters and representative compounds are selected from each cluster. This paper proposes the use subtractive clustering in compound clustering by finding the optimal data points to be defined as a cluster centers based on the density of surrounding data points. The technique resolves the problem of determining the suitable number of clusters for the data. Different values of cluster radii and inter-cluster squash factor have been evaluated. For subtractive clustering, good values of squash factor are between 0.375 and 0.45 and the cluster radii from 0.35 to 0.45 because they always give the highest proportion of active structures in active cluster datasets. The results obtained from subtractive clustering has also been used in fuzzy c-mean (FCM) and K-means. We found that the proportion of actives in active cluster subsets are better when FCM and K-means are based on the results produced by subtractive clustering compared to results from subtractive clustering. K-means produced the best results among the three clustering methods.

Key words: Subtractive clustering, cluster analysis, compound databases, compound clustering.

1.0 Introduction

The drug design technologies have already produced a tremendous amount of data that requires proper methods of data analyzing. The dramatic increase of resulting compound data has encouraged researchers in the field to look at ways of applying various machine learning techniques for data analysis. In the early stages of a drug discovery project, the emphasis is on lead generation process, in which an attempt is made to optimize the molecular diversity of the initial library produced for compound screening. Due to the similar property principle [1], structurally similar compounds can be expected to exhibit similar properties and biological activities. It is thus undesirable to test a large number of structurally similar compounds for many reasons. Maximizing the diversity of a subset is assumed to enhance the chances of finding active compounds of various structural types in screening experiments [2].

There are many approaches for compound selection such as cluster-based compound selection, dissimilarity-based compound selection, partition-based compound selection and optimization-based compound selection [2]. Among these different approaches, cluster-based or clustering has become the most commonly used in compound selection. Clustering is an unsupervised learning problem, where only inputs are available and no target outputs are

[*] *Corresponding author: Email: naomie@fsksm.utm.my*

predefined by the users. 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. Apart from compound selection, compound clustering can also be used to predict certain properties of the chemical compounds by looking at properties of compounds in the same clusters and summarizing contents of chemical databases [3].

In this work, we study the performance of subtractive clustering for clustering of chemical compounds and compare the result from subtractive clustering with fuzzy c-means (FCM) and K-means clustering.

2.0 Experimental Design

2.1 Dataset

This experiment uses 1000 compounds from the MDL Drug Data Report (MDDR) database, containing molecules of drugs launched or under development, as referenced in the patent literature, conference proceedings, and other sources [5]. These molecules were represented by topological indices generated using the Dragon software. The 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 highly discriminative power and so are able to distinguish slight variations in molecular structure. 99 topological indices which includes Zagreb index, quadratic index, Narumi simple topological index, total structure connectivity index, Wiener index and Balaban index have been used in the experiment. For every bioactivity considered, a particular compound can be regarded as either active or inactive. The effectiveness of the clusters produced will be tested based on the clusters ability to separate actives and inactive compound into different set of clusters.

Although subtractive clustering will determine the clusters to be produced, the radius of the cluster is critical for the subtractive clustering to work effectively. To estimate the optimum values for the radius *ra* of the clusters, we have used some training data to learn the optimum value of radius *ra* based on the proportion actives inside a clusters (*Pa*). We used a 5-fold cross-validation in which the dataset is partitioned into 5 subsets to estimate the optimum value for the radius *ra*.

2.2 Algorithms

2.2.1 Subtractive clustering

Subtractive clustering operates by finding the optimal data point to be defined as a cluster center, based on the density of surrounding data points. All data points within the radius distance of these points are then removed, in order to determine the next data cluster and its center. This process is repeated until all of the data is within the radius distance of a cluster center. To avoid obtaining closely spaced cluster centers, we set *rb* to be somewhat greater than *ra*.

Consider a collection of *n* data points *(x1,x2 ...,xn)* in an *m* dimensional space. In our case of topological indices, we have normalized the topological values in each dimension so that their coordinate ranges in each dimension are equal; i.e., the data points are bounded by a hypercube. We consider each data point as a potential cluster center and define a measure of the potential of data point *xi* as

$$P_i = \sum_{j=1}^{n} \exp(-\alpha \| x_i - x_j \|^2) \quad \text{----------------- (1)}$$

, where $\alpha = \frac{4}{r_a^2}$

P_i = Potential value of data point i
xi = *ith* data points
n = total number data points
exp = exponent
ra = Radii or radius defining a neighborhood

Thus, the measure of potential for a data point is a function of its distances to all other data points [6]. A data point with many neighboring data points will have a high potential value. The constant *ra* is the effective radius defining a neighborhood; data points outside this radius has little influence on the potential. Using the square of the distance eliminates the square root operation that otherwise would be needed to determine the distance itself. After the potential of every data point has been computed, we select the data point with the highest potential as the first cluster center. Let $x_1{}^*$ be the location of the first cluster center and $P_1{}^*$ be its potential value. We then revise the potential of each data point *xi* by the formula

$$P_i = P_i - P_1{}^* \exp(-\beta \| x_i - x_1{}^* \|^2) \quad \text{----------- (2)}$$

where

$$\beta = \frac{4}{r_b^2}$$

rb = Squash value
$x_1{}^*$ = first cluster center point

Thus, we subtract an amount of potential from each data point as a function of its distance from the first cluster center. The data points near the first cluster center will have greatly reduced potential, and therefore will unlikely be selected as the next cluster center. The constant *rb* is the effective radius to define the neighborhood which will have measurable reductions in potential. To avoid obtaining closely spaced cluster centers, we set *rb* to be somewhat greater than *ra* ; a good choice is *rb* = 1.5 *ra* [6].

When the potential of all data points has been revised according to Eq. (2). We select the data point with the highest remaining potential as the second cluster center. We then further reduce the potential of each data point according to its distance to the second cluster center. In general, after the *k'th* cluster center has been obtained, we revise the potential of each data point by the formula

$$P_i = P_i - P_k{}^* \exp(-\beta \|x_i - x_k{}^*\|^2) \text{ ------ } (3)$$

$x_k{}^*$ = k'th cluster center point
$P_k{}^*$ = Potential value of $x_k{}^*$

where $x_k{}^*$ is the location of the *k'th* cluster center and $P_k{}^*$ is its potential value. The process of acquiring new cluster centers and revising potentials repeats until the potential value is below the acceptance value deemed as important in affecting the final clustering results.

The algorithm for subtractive clustering is as below

```
START
    Step 1 : Select the parameters values:
                $ra$ : radius
                $rb$ : Squash factor
                $\bar{\varepsilon}$ : Accept Ratio
                $\underline{\varepsilon}$ : Reject Ratio
    Step 2  : Normalize the data into a unit hyperbox
    Step 3 : Compute the initial potentials for each data points using Eq. 1
    Step 4: Find the data point with highest potential value to be the first cluster center $P_1{}^*$.
    Step 5 : Revise the potential of all data points using  Eq. 3 until criteria below:

        if $P_k{}^* > \bar{\varepsilon}\ P_1{}^*$ ;
            Accept $x_k{}^*$  as a cluster center and continue.
        else if $P_k{}^* < \underline{\varepsilon}\ P_1{}^*$ ;
            Reject $x_k{}^*$ and end the clustering process
        else
                Let dmin = shortest of the distances between $x_k{}^*$ and all previously found
                cluster centers.

            If $\frac{d\min}{r_a} + \frac{p_k{}^*}{p_1{}^*} \ge 1$
                Accept $x_k{}^*$ as a cluster center and continue.
            else
                Reject $x_k{}^*$ and set the potential at $x_k{}^*$ to 0.
                  Select the data point with the next highest potential as the new $x_k{}^*$, re-
                  test and calculate its potential value using Eq. 3.
            endif
        endif
```

Here $\bar{\varepsilon}$ specifies a threshold for the potential above which we will definitely accept the data point as a cluster center; $\underline{\varepsilon}$ specifies a threshold below which we will definitely reject the data point. We use $\bar{\varepsilon}$ = 0.5 and $\underline{\varepsilon}$ = 0.15. If the potential falls in the gray region, we check if the data point provides a good trade-off between having a reasonable potential and being sufficiently far from existing cluster centers [6].

3.0 Results and Discussion

The result from the subtractive clustering is evaluated based on their ability to separate active/inactive structures into different clusters. This criterion will allow sampling of the range of activities in the datasets and minimize the chances that any activity is missed when an inactive compounds is selected as the representative of a cluster containing actives [2]. The more active structures are in a cluster, the higher possibility that an active structure will be selected as a representative for further analysis. For both of the analyses, different cluster radius in the range of 0.2 to 0.5 and squash factor in the range of 0.3 to 0.75 are used. This is done to see the effect of different radius cluster (*ra*) and squash factor (*rb*) to the clusters produced.

Figure 1 shows the comparison of proportion of active structures in active clusters, defined as clusters with at least one active structure, with different cluster radius from 0.2 to 0.5 when different training data sets are used. Radius value from 0.2 to 0.5 has been chosen because good values for radii are usually between 0.2 and 0.5 [6] and small radii values generally result in a few large clusters. From Figure 1, we can see that the highest proportion of actives (*Pa)* are obtained with cluster radius from 0.35 to 0.45. Thus, we can conclude that the best cluster radius is from 0.35 to 0.45 in modeling subtractive clustering for chemical compound clustering.

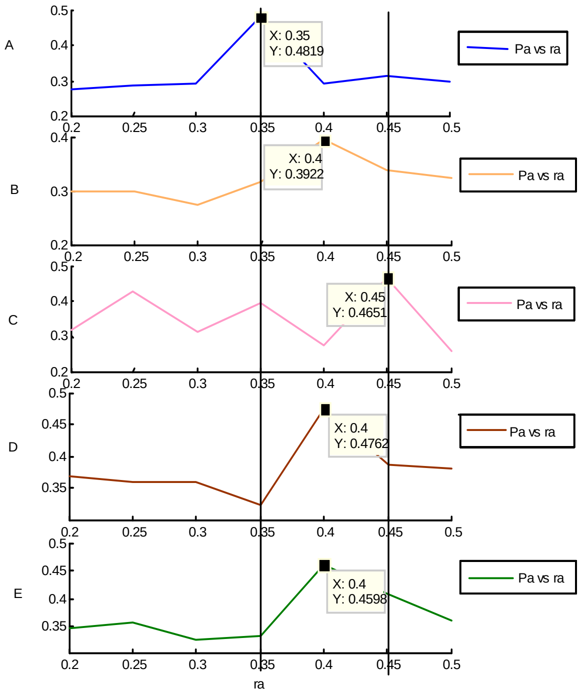

Figure 1 Results from subtractive clustering based on their proportion of actives in active clusters (Pa) for 5 experiments.

Another analysis is based on the squash factor used in subtractive clustering. Figure 2 shows the proportion of active structures which have been produced using squash factor from 0.3 to 0.75. The reason that we chose the rank between 0.3 and 0.75 is to avoid obtaining closely spaced cluster centers. We set *rb* to be somewhat greater than *ra* and the good choice is *rb* ≈ 1.5 *ra* [6]. From the graph, we can conclude that good values of squash factor are between 0.375 and 0.45 for subtractive clustering in training data sets.

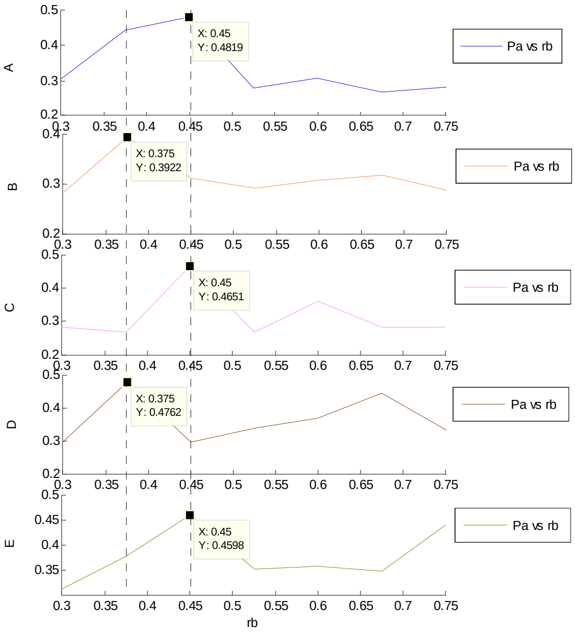

Figure 2 Results from subtractive clustering based on their proportion of actives in active clusters (Pa) for 5 experiments.

From the results, it is clear that choosing very small *ra* or very large *ra* will result in poor accuracy because when a very small *ra* is chosen the density function will not take into account the effect of neighboring data points; while if a very large *ra* is taken, the density function will include most data points in the data space. The squash factor *rb* is used to determine the neighborhood of a cluster center within which the existence of other cluster centers is discouraged so as to quash the potential for outlying points is to be considered as part of that cluster. This is the reason that the number clusters produced decreased when the squash factor increased although the same radii have been used in model.

The results obtained from subtractive clustering have also been used in fuzzy c-mean (FCM) and K-means. The number cluster produced from subtractive will be used in FCM with fuzziness index = 2.0 and no. iteration = 100. The reason for choosing the fuzziness index = 2 is because it has been found that FCM produced high proportion of actives using this value [7]. For K-means method, again the number cluster produced from subtractive clustering is used. Subtractive clustering is calculated only at every data point with the difference of a density function, instead of at every grid point. So the data points themselves are the candidates for cluster centers. This will reduce the number of computations significantly, and making it linearly proportional to the number of input data instead of being exponentially proportional to its dimension. Although the ra and rb of subtractive clustering are optimized from training data sets but the Pa proceeded from FCM is still higher than subtractive clustering. From Figure 3, we found that the proportion of actives in active cluster subsets are better when FCM and K-means are based on the results produced by subtractive clustering compared to the use of subtractive clustering by itself. K-means produced the best results among the three clustering methods.

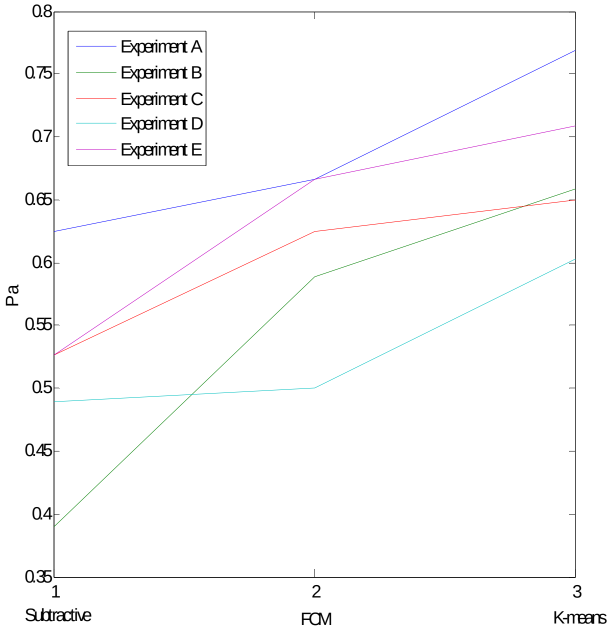

Figure 3 Results of comparison based on Proportion of actives in active clusters (Pa) for 5 experiments using different data sets.

K-means clustering works on finding the cluster centers by trying to minimize a cost function. It alternates between updating the membership matrix and updating the cluster centers respectively, until no further improvement in the cost function is noticed. Since the algorithm initializes the cluster centers randomly, its performance is affected by those initial cluster centers. It is therefore understandable that a good approximation of the number of clusters can improve the results.

4.0 Conclusion

The study presents the potential use of subtractive clustering for clustering chemical compound databases. Based on our results, the cluster radius to use for optimum separation between actives and inactives are between 0.35 with 0.45 whilst good squash factor are between 0.375 with 0.45. We have also shown that the results are even better if the number of clusters obtained is used for FCM and K-means clustering. However, we are yet to combine the idea from subtractive clustering directly into FCM and K-means algorithms.

For this study, the dataset used for the experiment was represented by topological descriptors. Experiments should also be conducted using other descriptors to see if the possibility of using subtractive clustering for other molecular descriptors. Apart from subtractive clustering, other density search clustering methods approaches can also be used, such as the Mountain clustering and the Taxmap methods.

References

[1] Johnson, M.A. and Maggiora, G.M. 1990. *Concepts and Application of Molecular Similarity*. New York: John Wiley and Sons.

[2] Holliday, J.D., Salim, N. and Willett, P. 2005. On The Magnitudes of Coefficient Values in The Calculation of Chemical Similarity and Dissimilarity. In Lavine, B.K. (Ed.) *Chemometrics and Chemoinformatics*. Washington, D.C. : American Chemical Society.

[3] Brown, R.D., Bures, M.G. and Martin, Y.C. 1995. Similarity and cluster analysis applied to molecular diversity. *American Chemical Society Meeting*. 209:3-COMP. Anaheim, California.

[4] Hecht, P. 2002. High-throughput screening: beating the odds with informatics-driven chemistry. *Current Drug Discovery*. January 2002: 21-24.

[5] MDL's Drug Data Report: http://www.mdli.com/products/knowledge/drug_data_report/index.jsp

[6] S.L.Chiu. 1994. Fuzzy model identification based on cluster estimation. *Journal of Intelligent Fuzzy Systems.* 2:267-278.

[7] Huspi,S.H. and Salim,N. 2005. Cluster-Based Compound Selection for Bioactivity Testing Using Fuzzy Clustering Approach. *Journal of Advancing Information Management Studies*. 2(1):31-45.