Fuzzy C-means method for clustering microarray data
Doulaye Dembélé* and Philippe Kastner

Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS-IMSERM-ULP,
BP 10142, 67404 Illkirch Cedex, France

Received on August 14, 2002; revised on November 14, 2002; accepted on January 3, 2003

ABSTRACT
Motivation: Clustering analysis of data from DNA microarray hybridization studies is essential for identifying biologically relevant groups of genes. Partitional clustering methods such as K-means or self-organizing maps assign each gene to a single cluster. However, these methods do not provide information about the influence of a given gene for the overall shape of clusters. Here we apply a fuzzy partitioning method, Fuzzy C-means (FCM), to attribute cluster membership values to genes.

Results: A major problem in applying the FCM method for clustering microarray data is the choice of the fuzziness parameter \( m \). We show that the commonly used value \( m = 2 \) is not appropriate for some data sets, and that optimal values for \( m \) vary widely from one data set to another. We propose an empirical method, based on the distribution of distances between genes in a given data set, to determine an adequate value for \( m \). By setting threshold levels for the membership values, genes which are tightly associated to a given cluster can be selected. Using a yeast cell cycle data set as an example, we show that this selection increases the overall biological significance of the genes within the cluster.

Availability: Supplementary text and Matlab functions are available at http://www-igbmc.u-strasbg.fr/fcm/

Contact: doulaye@titus.u-strasbg.fr

1 INTRODUCTION
DNA microarrays are used to monitor genes expression in many areas of biomedical research. To analyze the increasing amount of data produced by this technology, clustering has become inevitable. Clustering methods can be roughly divided into two groups: hierarchical and partitional methods (Theodoridis and Koutroumbas, 1999). The results of hierarchical methods are represented as dendrograms, each branch representing a group of genes with similar behavior (Eisen et al., 1998). Hierarchical methods suffer however from the non-uniqueness of the dendrogram (Morgan and Ray, 1995).

Partitional methods aim to find the best partition of genes into \( K \) clusters in such a way that one criterion (e.g. the total inertia of clusters) is optimized. Here we use the Fuzzy C-means (FCM) algorithm (Bezdek, 1981). In addition to the specification of the number \( K \) of clusters in the data set, the FCM method requires to choose \( m \), the fuzziness parameter. There is little literature on the choice of this parameter (Bezdek, 1981; McBratney and Moore, 1985). Here we propose a new method which allows to compute an upper bound value for \( m \). We then choose \( m \) independently of \( K \). We show that FCM clustering of microarray data, combined with threshold-based gene selection, offers a convenient way of defining subsets of genes which are more tightly associated to a given cluster.

2 MATERIALS AND METHOD

2.1 Data sets
2.1.1 Serum data. This data set is described and used in (Iyer et al., 1999). It can be downloaded from: http://www.sciencemag.org/feature/data/984559.shl and corresponds to the selection of 517 genes whose expression vary in response to serum concentration in human fibroblasts.

2.1.2 Yeast data. In this data set, the expression profiles of 6200 yeast genes were measured every 10 min during two cell cycles in 17 hybridization experiments (Cho et al., 1998). We used the same selection of 2945 genes made by Tavazoie et al. (1999). In that selection, the data for the time points 90 and 100 min are excluded.

2.1.3 Human cancer data. This data set represents gene expression patterns of 9703 genes in 60 human cancer cell lines and can be downloaded from http://discover.nci.nih.gov/nature2000/. The complete data set contains missing values. We first selected genes which have at most three missing values. There were 8161 such genes. We used the k-nearest neighbors (k-NN) impute algorithm described in (Troyanskaya et al., 2001) to estimate the missing values, then selected genes with expression levels that varied by at least fold in at least four...
of the 60 cell lines. This filter gave the 728 genes used in this study.

All data were normalized in such a way that every gene had an average expression value of zero and a standard deviation equal to 1. We note \( X = \{ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \} \), where \( \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{T} \) is a \( p \)-dimensional vector representing gene \( i \) with its \( p \) experiments, the input data set for the clustering method.

### 2.2 Algorithms

The fuzzy clustering algorithm links each gene to all clusters via a real-valued vector of indexes. The values \( u_{ki} \) of the components of this vector lie between 0 and 1. For a given gene, an index close to 1 indicates a strong association to the cluster. Inversely, indexes close to 0 indicate the absence of a strong association to the corresponding cluster. The vector of indexes defines thus the membership of a gene with respect to the various clusters. Membership vector values \( u_{ki} \) and cluster centroids \( \mathbf{c}_k \) can be obtained after minimization of the total inertia criterion (Bezdek, 1981):

\[
J(K, m) = \sum_{k=1}^{K} \sum_{i=1}^{N} (u_{ki})^m d^2(\mathbf{x}_i, \mathbf{c}_k) \tag{1}
\]

\[
d^2(\mathbf{x}_i, \mathbf{c}_k) = (\mathbf{x}_i - \mathbf{c}_k)^T A_k (\mathbf{x}_i - \mathbf{c}_k) \tag{2}
\]

where \( 1 \leq i \leq N \) and \( 1 \leq k \leq K \).

In Equation (1), \( K \) and \( N \) are respectively the number of clusters and the number of samples (or genes) in the data, \( m \) is a real-valued number which controls the ‘fuzziness’ of the resulting clusters, \( u_{ki} \) is the degree of membership of gene \( \mathbf{x}_i \) in cluster \( k \), and \( d^2(\mathbf{x}_i, \mathbf{c}_k) \) is the square of distance from gene \( \mathbf{x}_i \) to centroid \( \mathbf{c}_k \). In Equation (2), \( A_k \) is a symmetric and positive definite matrix.

Equation (3) indicates that empty clusters are not allowed. Scalar \( m \) is any real-valued number greater than 1 (see below for its choice). When matrix \( A_k \) is the identity matrix, then \( d^2(\mathbf{x}_i, \mathbf{c}_k) \) corresponds to the square of the Euclidian distance. From Equation (1), parameters of interest are the cluster centroid vectors \( \mathbf{c}_k \) and the components of the membership vectors \( u_{ki} \). These unknown parameters can be obtained using the following algorithm (Bezdek, 1981):

1. 
(0) Initialization: Fix \( K, m \) and choose any product norm metric for calculation of \( d^2(\mathbf{x}_i, \mathbf{c}_k) \). Select randomly \( K \) samples as initial centroids \( \mathbf{c}_k^{(0)} \) and then form partitions of all others samples around these centroids to obtain the initial partition matrix \( U^{(0)} = [u_{ki}], k = 1, \ldots, K \) and \( i = 1, \ldots, N \). At step \( l, l = 1, 2, \ldots, \) perform the following steps:

   (1) Computation of centroids \( \mathbf{c}_k^{(l)} \):

\[
\mathbf{c}_k^{(l)} = \frac{\sum_{i=1}^{N} (u_{ki}^{(l-1)})^m \mathbf{x}_i}{\sum_{i=1}^{N} (u_{ki}^{(l-1)})^m}; k = 1, 2, \ldots, K \tag{4}
\]

   (2) Computation of membership values \( u_{ki}^{(l)} \):

\[
u_{ki}^{(l)} = \frac{1}{\sum_{i=1}^{K} \left( \frac{d^2(\mathbf{x}_i, \mathbf{c}_k^{(l)})}{d^2(\mathbf{x}_i, \mathbf{c}_i^{(l)})} \right)^{1/(m-1)}} \left( \frac{1}{\sum_{i=1}^{K} \left( \frac{d^2(\mathbf{x}_i, \mathbf{c}_k^{(l)})}{d^2(\mathbf{x}_i, \mathbf{c}_i^{(l)})} \right)^{1/(m-1)}} \right)_{\forall i \in I_l} \tag{5}
\]

   (3) Repeat (1) and (2) until stabilization, i.e. \( \|U^{(l)} - U^{(l-1)}\| < \epsilon, l > 1 \)

After several passes through (4) and (5), the algorithm will stop, i.e. the error between two consecutive values of the constrained fuzzy partition matrix \( U \) will be smaller than an \( a \ priori \) specified level. Convergence of FCM has been proven (Bezdek, 1981, pp. 80–86).

2.2.1 The fuzziness parameter

2.2.1.1 \( m \) is not an appropriate fuzziness parameter for microarray data. Once the norm metric for calculating \( d^2(\mathbf{x}_i, \mathbf{c}_k) \) is chosen, the minimization of criterion \( J(K, m) \) given in relation (1) depends on the choice of \( m \). In the literature about FCM, \( m \) is commonly fixed to 2. This choice allows easy computation of the membership values \( u_{ki} \). However, when we used \( m = 2 \) for the microarray data sets, we observed that in two cases (yeast and cancer data sets), all the membership values were similar. Thus in these cases, FCM failed to extract any clustering structure. In the case of the serum data set, a clustering structure was found, but all membership values were relatively low (Fig. 1a), suggesting that this FCM setting failed to tightly associate any gene to any cluster. These observations prompted us to search for a method to evaluate appropriate values for \( m \).

2.2.1.2 Upper bound value for \( m \). It was shown (Bezdek, 1981, p. 73) that when \( m \) goes to infinity, values of \( u_{ki} \) go to \( \frac{1}{K} \). Thus, for a given data set, there is an upper bound value for \( m \) (\( m_{ub} \)), above which the membership values resulting from FCM are equal to \( \frac{1}{K} \). As a first step towards the evaluation of an appropriate value for \( m \), we first attempted to estimate \( m_{ub} \). From Equation (5),
we note that membership values $u_{ki}$ depend on distances between genes and cluster centroids. For complex data sets, it is reasonable to make the approximation that the cluster centroids will be close to some genes. We thus made the hypothesis that when $m$ varies, there might be a relationship between the FCM membership values and the coefficient of variation (cv) of the set of distances between genes:

$$Y_m = \left[ \{ d^2(x_i, x_k) \}^{\frac{1}{m-1}} \right] ; \quad k \neq i, 1, 2, \ldots, N$$  \hspace{1cm} (6)

Note that $Y_m$ depends only on the initial data set and $m$, and is thus completely independent of the FCM results.

To test the above hypothesis, we used several data sets of known structure (e.g. the iris data set, downloadable from ftp://ftp.ics.uci.edu/pub/machine-learling-databases/ and two synthetic data sets generated for the purpose of this study; see supplementary text). For each data set, we varied $m$ and determined the cv of $Y_m$. We also ran the FCM algorithm to determine the distribution of the membership values. In each case, we observed that the values of $m$ which leads to membership values close to $\bar{Y}$ gave a cv of $Y_m$ close to $0.03p$, $p$ being the data dimension. We have no theoretical justification for this observation. We propose to use it to solve the following equation to evaluate $m_{ub}$:

$$cv[Y_m] = \frac{\sigma_{Y_m}}{\bar{Y}_m} \approx 0.03 p$$  \hspace{1cm} (7)

where $\sigma_{Y_m}$ and $\bar{Y}_m$ are respectively the standard deviation and the mean of the set $Y_m$.

We solved Equation (7) numerically by using a dichotomy search strategy. Initially we set $m = 2$ and computed $cv[Y_2]$. This value allowed us to decide the direction of search: in $[1, 2]$ if $cv[Y_2] < 0.03p$, in $[2, \infty]$ if $cv[Y_2] > 0.03p$ and $m_{ub} = 2$ if $cv[Y_2] \approx 0.03p$. If $m_{ub}$ was not equal to 2, we performed successive choices of $m$ in the correct direction and computed $cv[Y_m]$ until $cv[Y_m] \approx 0.03p$. Column 4 of Table 1 gives the upper bound value for the data sets used after application of the algorithm proposed. For each microarray data set, we confirmed that the use of $m_{ub}$ as fuzziness parameter leads to FCM membership values close to $\bar{Y}$ (Fig. 1b).

### Table 1. Parameters used for the FCM algorithm

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>$p$</th>
<th>$m_{ub}$ exper.</th>
<th>$m_{ub}$ rand.</th>
<th>$m$ used</th>
<th>K used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum</td>
<td>517</td>
<td>13</td>
<td>2.548</td>
<td>1.92</td>
<td>1.25</td>
<td>10</td>
</tr>
<tr>
<td>Yeast</td>
<td>2945</td>
<td>16</td>
<td>1.71</td>
<td>1.84</td>
<td>1.17</td>
<td>16</td>
</tr>
<tr>
<td>Cancer</td>
<td>728</td>
<td>60</td>
<td>1.2</td>
<td>1.17</td>
<td>1.112</td>
<td>20</td>
</tr>
</tbody>
</table>

**Fig. 1.** Influence of the fuzziness parameter $m$ on the distribution of membership values. Boxplot representations (Hoaglin et al., 2000) of sorted membership values from FCM clustering are shown. For fixed values of $m$, the $K$ membership values of each gene were sorted in decreasing order. For a point in each plot, horizontal segments are 99 centile, third quartile, median, first quartile and first centile values respectively; isolated segments represent outliers. (a) Distribution of membership values when $m$ is fixed to 2. Note that, in the case of the yeast and cancer data sets, all membership values for all genes were equal to $\frac{1}{K}$. (b) Distribution of membership values when $m$ is equal to the computed upper bound value $m_{ub}$. (c) Effect of varying $m$ in the case of the serum data.
to only one cluster, and the clustering is similar to that obtained with K-means. Thus, the selected value for \( m \) appears to be a good compromise between the need to assign most genes to a given cluster, and the need to discriminate genes that classify poorly.

2.2.2 Running the FCM algorithm. For one run of the algorithm in Equations (4) and (5), we selected randomly \( K \) genes as initial centroids. To avoid the algorithm taking too long before convergence, we used two stopping criteria: the error between two consecutive values of the constrained fuzzy partition matrix and a pre-specified maximum number of iterations. The minimum error is chosen equal to 0.001 which is compared at each iteration to the Frobenius norm \(^1\) of the matrix that results from the difference between partition matrices at the present iteration and the previous one. The maximum number of iterations was fixed to 500. We performed 30 runs of the FCM algorithm, then we kept the membership values and centroids which correspond to the lowest value of the criterion \( J(K, m) \). One run of the FCM algorithm leads to a local solution. Using many runs with different initializations allows to explore the entire data space for the best solution.

2.2.2.1 The number \( K \) of clusters. We have used the CLuster Identification via Connectivity Kernels (CLICK) algorithm (Sharan and Shamir, 2000) to have an estimation of the number of clusters. CLICK combines graph-theoretic and statistical techniques for automatic identification of clusters in a data set.

2.2.2.2 Assessment of the quality of the clusters with a silhouette measure. To assess the quality of clusters, we used the silhouette measure proposed by Rousseeuw (1987). For computing the silhouette value of a gene \( x_i \), we first estimate two scalars \( a(x_i) \) and \( b(x_i) \). Let us note \( C_r \) the cluster to which gene \( x_i \) belongs. The scalar \( a(x_i) \) is the average distance between gene \( x_i \) and all other genes of \( C_r \). For any other cluster \( C_s \neq C_r \), let \( d(x_i, C_s) \) denote the average distance of gene \( x_i \) to all genes of \( C_s \). The scalar \( b(x_i) \) is the smallest of these \( d(x_i, C_s), r \neq s = 1, \ldots, K \). The silhouette \( s(x_i) \) of gene \( x_i \) is the ratio \( \frac{\max(a(r(i)), b(i)) - a(i)}{\max(a(r(i)), b(i))} \). The silhouette value lies between \(-1\) and \(1\). When its value is less than zero, the corresponding gene is poorly classified.

3 RESULTS

The values of \( m \) and \( K \) used for FCM clustering of the three data sets are given in Table 1. The values for \( m_{ab} \) differ markedly from one data set to another, being in particular low for the cancer data set. Therefore no general assumption can be made for values of \( m \) that might give good FCM results with microarray data sets. Evaluation of \( m_{ab} \) prior to clustering thus greatly facilitates the choice of an appropriate value for \( m \).

For each data set, we also ran the FCM algorithm for the randomized data sets (Table 1 and Fig. 2). A clustering structure could be found for the randomized serum and yeast data, but not the cancer data, probably owing to the lower complexity of the former data sets.

3.1 Distribution of membership values

Figure 2 shows the distribution of the sorted membership values ranked in decreasing order for the real data and a representative randomized data set. In Figure 3, we show

\(^1\)the Frobenius norm of a matrix \( U \) of sizes (\( KN \)) is defined as: \[ \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{N} (U(i, j))^2} \]
scatterplots representing the highest (horizontal axis) and second highest (vertical axis) $u_{ki}$ values for each gene.

From these Figures, two observations can be made:

(i) The distribution of the membership values varies markedly from one data set to another. For the serum data, the behavior of each gene can be almost entirely determined by its first and second membership values. In contrast, in the case of the cancer data set, the dispersion of the $u_{ki}$ values was higher, and a significant proportion of the genes ($\approx 35\%$) exhibited only loose association to any cluster ($\max of u_{ki} < 0.5$). For both the yeast and cancer data, the fraction of the genes for which ($\max of u_{ki} + (2\text{nd }max of u_{ki}) \leq 0.9$ was markedly higher than for the serum data set (genes located on the left of the oblique lines in Fig. 3).

(ii) Relatively high membership values can be obtained for randomized data sets, most notably with the yeast data set. In contrast, only marginal clustering structure could be found in the randomized serum data. This observation shows that some data sets are intrinsically more noisy than others, and that a clustering structure can be found even in data sets which do not have any biological significance.

3.2 Selection of genes tightly associated to clusters

The above observations suggest that it may be of interest to restrict clustering only to genes which display a strong association to a given cluster.

In the examples shown in Figures 4 and 5, we chose as threshold the median of the highest membership values of genes. This threshold is shown by the vertical lines in Figure 3.
in Figures 3 and 4a. In the case of the cancer data set, the expression profile of the selected genes shows well separated and homogeneous clusters (Fig. 4c). In contrast, the expression profile of the excluded genes (delimited by the left side of vertical line in Figure 4a) exhibited only poor clustering structure (Fig. 4b).

To further confirm that gene selection based on threshold values for uu_kl leads to clusters which are better separated, we calculated the silhouette values for each gene within each cluster, either taking all genes into account, or only the genes selected as having a (max of uu_kl) higher than the selected threshold. Boxplots of the silhouette values are shown for each cluster in Figure 5. As expected, threshold based selection of genes increases the overall silhouette value for the cluster. Significantly, the effect of gene selection was more pronounced for some clusters. For instance, in the case of the serum data, the quality of cluster 8 (which contains many genes with negative silhouette values without gene selection) was clearly improved. In the case of the cancer data, cluster 4 (which had a negative mean silhouette value prior to selection) became one of the best separated clusters after selection (Fig. 5c). In contrast, the clusters which were already well separated prior to gene selection (such as cancer clusters 7 and 9) gained relatively little from the process.

In the case of the yeast data, Tavazoie et al. (1999) have shown a significant enrichment for genes belonging to given functional categories in some of the clusters obtained with the K-means method. It is likely that most of the genes belonging to these enriched functional categories are biologically significant members of the corresponding clusters. Similarly to Tavazoie et al. (1999), we found in six distinct ‘raw’ clusters resulting from FCM, 16 groups of genes corresponding to enriched functional categories (Table 2). We then applied a threshold-based selection to shed the most loosely associated genes from each cluster (only genes for which the highest membership value was above 0.54 were selected). For each of the 16 groups of genes belonging to enriched functional categories, we determined the percentage of genes retained in the corresponding restricted cluster, and compared it to the overall percentage of genes retained for that cluster (Table 2). Remarkably, in most cases, the genes from these groups were retained more frequently than the average genes in the clusters. This increase in the degree of functional category enrichment was in some cases highly significant (‘ribosome biogenesis’ and ‘organization of cytoplasm’ categories in cluster 11 or ‘organization of nucleus’ category in cluster 16; Table 2). In addition, even though the increased enrichment seen for some of the functional categories was, when considered in isolation, not very significant, the fact that we never observed a decrease in the level of enrichment for any of the 16 groups of genes evaluated, is by itself highly significant (p < 10^-4). This example shows that selecting genes which are the most tightly associated to clusters increases the biological relevance of the genes within the clusters.

4 DISCUSSION

Even though clustering with fuzzy methods had been proposed previously for gene expression data (Dougherty et al., 2002), the use of this method has been hindered so far by problems associated with the choice of m. We propose a novel method which computes an upper bound value for m, which is then used to choose m independently of the number K of clusters before applying the clustering algorithm. The computation of m_ub uses only first and second-order sample statistics (mean and
between genes are on average smaller for...show coherent behavior within clusters. In the case of the...will assign to each exhibiting tight association to given clusters. Conventional partitional clustering methods force all genes into clusters, even those for which the variations in expression do not fit...FCM clustering might also be useful to unravel complex modes of regulation for some genes. It is indeed known that given genes are subject to regulation by several molecular pathways. The overall expression pattern for a given gene may therefore correspond to the superimposition of distinct patterns, each corresponding to a given mode of regulation. In this respect, the clusters defined by the secondary or third highest membership value might identify such secondary modes of regulation. However, in the data sets used for the present analysis, the second membership value usually identified clusters that had profiles related to those identified by the highest $M_k$ value. It is likely that the data sets used here are too simple in their structure (e.g. kinetics or cell lines representing only a few distinct cell types in the cancer data) to allow superimposition of entirely distinct modes of regulation. We anticipate however, that fuzzy clustering might become a useful tool to dissect the various regulatory pathways that control the expression pattern of a given gene when dealing with highly complex data sets.

**ACKNOWLEDGEMENTS**

We thank W. Raffelsberger for critical reading of the manuscript. This work was supported by the Institut National de la Santé et de la Recherche Médicale, the
Centre National de la Recherche Scientifique, the Hôpital Universitaire de Strasbourg and the Centre National de Recherche en Génomique. D.D. is recipient of a post-doctoral fellowship from the Groupement d’interêt public—Hoechst-Marion-Roussel.

REFERENCES


