Fuzzy Clustering of Gene Expression Data

Matthias E. Futschik and Nikola K. Kasabov
Department of Information Science, University of Otago
P.O. Box 56, Dunedin, New Zealand
email: mfutschik@infoscience.otago.ac.nz, nkasabov@otago.ac.nz

Abstract– Microarray techniques have recently made it possible to monitor simultaneously the activity of thousands of genes. They offer new insights into the biology of a cell. However, the data produced by microarrays poses several challenges to overcome. One major task in the analysis of microarray data is to reveal structures in the data despite its large noise component. We used fuzzy c-means (FCM) clustering in this study to achieve a robust analysis of gene expression time-series. We address the issues of parameter selection and cluster validity. Using statistical models to simulate gene expression data, we show that FCM can detect genes belonging to different classes. This may open the way for the study of fine-structures in microarray data.

I. Introduction

By enabling researchers the simultaneous measurement of the activity of many thousands genes, microarrays have revolutionized the study of complex genetic networks. They have become very powerful techniques in the systematic study of gene regulation. A landmark experiment was the study of the yeast cell cycle using microarrays that contained every gene of the yeast genome [1] delivering an unexpected richness of patterns of gene activities in the cell. To reveal these structures, a first step in the data analysis is frequently the application of clustering methods. One of the main purposes for cluster analysis of gene expression data is to infer the function of novel genes by grouping them with genes of well-known functionality. This is based on the observation that genes which show similar activity patterns (*coexpressed genes*) are often functionally related and are controlled by the same mechanisms of regulation (*coregulated genes*). The gene clusters generated by cluster analysis often relate to certain functions e.g. DNA replication, or protein synthesis. If a novel gene of unknown function falls into such a cluster, it is likely that this gene serves the same function as the other members of this cluster. This 'guilt-by-association' method makes it possible to assign functions to a large number of novel genes by finding groups of co-expressed genes across a microarray experiment [2].

Different cluster algorithms have been applied to the analysis of gene expression data: k-means, SOM and hierarchical clustering to name just a few [3]-[5]. They all assign genes to clusters based on the similarity of their activity patterns. Genes with similar activity patterns should be grouped together, while genes with different activation patterns should be placed in distinct clusters. The cluster methods used so far have been restricted to a one-to-one mapping: one gene belongs to exactly one cluster. While this principle seems reasonable in many fields of cluster analysis, it might be too limited for the study of microarray data. Genes can participate in different genetic networks and are frequently coordinated by a variety of regulatory mechanisms. For the analysis of microarray data, we may therefore expect that single genes can belong to several clusters. Several researchers have noted that genes were frequently highly correlated with multiple classes and that the definition of clear boarders between gene expression clusters seemed often arbitrary [2], [6]. This motivated us to use fuzzy c-means clustering (FCM) as a method that can assign single objects to several clusters.

A second reason for applying FCM clustering is the large noise component in microarray data due to biological and experimental factors. The activity of genes can show large variations under minor changes of the experimental conditions. Numerous steps in the experimental procedure contribute to additional noise and bias. An usual procedure to reduce the noise in microarray data is setting a threshold for a minimum variance of the abundance of a gene. Genes below this threshold are excluded from further analysis. However, the exact value of the threshold remains arbitrary due to the lack of an established error model and the use of filtering as preprocessing step may exclude interesting genes from further analysis. FCM may be a valuable approach because of its noise robustness.

A major problem to date for the critical assessment of any clustering approach in the field of gene expression data analysis is the missing of any benchmark data set for which the number of clusters is known. Different methods frequently yield different optimal numbers of clusters. Thus, a fair comparison between alternative clustering methods remains difficult using solely the original data. To achieve, nonetheless, a stringent comparison between different clustering algorithms, we introduce several statistical models for gene expression that are based on original data and use them for the generation of gene expression data with a controlled number of clusters. Additionally, by analyzing the performance of the clustering methods for data derived from different models, we gain insights in the applicability and limits of these clustering approaches.

In the next section we give first a brief introduction into the basic biology and the techniques used in a microarray experiment. This may facilitate the understanding of the data structure and the challenges posed to any clustering

0-7803-7280-8/02/$10.00 ©2002 IEEE

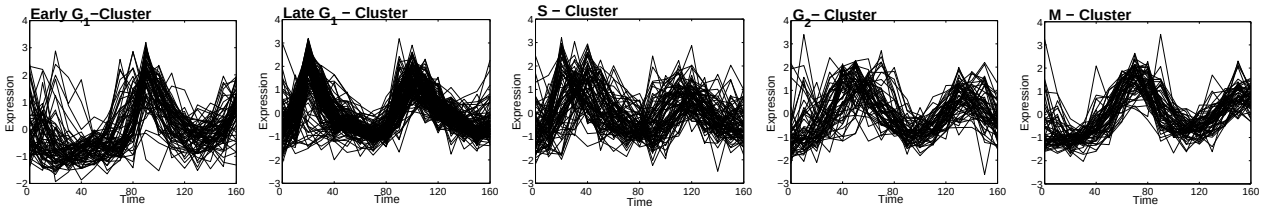

Fig. 1. Periodical clusters selected by Cho et al.

of microarray data. We describe further the data sets and the methods we used in this study. In the results section, we first focus on the parameter selection for FCM clustering. We then compare the noise robustness of FCM to hard k-means clustering. On model-based generated data sets, we show that FCM correctly assigns single genes to multiple clusters. Finally we discuss the results and describe future directions of our research.

II. Gene Expression Data

A. Biological background

Genes are the carrier of the information necessary for building and controlling of a cell as the basic unit of life. They are stored as DNA usually in the cell nucleus. The functions in a cell, however, are mainly performed by proteins. Both genes and proteins are linked together by RNA. RNAs are temporary copies of genes and constitute the intermediate step between the genes and their corresponding proteins. The process of the generation of the RNA and proteins is called the expression of a gene or *gene expression*. The amount of expressed RNA of a particular gene is generally an indicator how active this gene is. Highly active genes express more RNA than genes of low activity. Since every gene produces its own specific RNA, measuring the abundance of the RNA corresponding to each gene gives us an index for the activity of the genes. Microarray techniques enable us to monitor simultaneously the amount of RNA for thousands of genes. This is done by reverse transcription of RNA into cDNA (*copied DNA*). The generated cDNA is labelled by a dye and hybridized on gene specific probes on a slide. By measuring the amount of fluorescence evoked by a laser, the researcher can infer the abundance of the RNA for the screened genes in the sample. This is done for every sample in the experiment. The resulting data from the experiment is usually stored in a matrix form called the *gene expression matrix* G. Columns correspond to single samples and rows show the measured expression values for each genes used in the experiment. The collective expression values for a gene across samples constitute its *gene expression vector*.

B. Gene expression time series

We analyze in this study gene expression data derived from a time-series experiment of the yeast cell cycle [6]. Studying the yeast cell cycle is not only of biological importance, but also of medical interest. Many basic genetic mechanisms in yeast are similar in humans. Since the development of cancer is often linked to malfunctions in the cell cycle control, the analysis of the regulation mechanisms in yeast may reveal valuable information for medical researchers.

In the experiment by Cho et al., the expression values for (*Saccharomyces cerevisiae*) genes were recorded over a period of two cell cycles (160 min). The original data set consists of expression values for over 6000 genes taken over 17 evenly spread time points. In this study, however, we used a much smaller subset, since we seek to evaluate the performance of FCM clustering. For this task, we need prior knowledge of the data, especially the number of clusters we can expect. We studied therefore a set of genes that was selected by Cho et al. They selected by visual inspection 384 genes and categorized them into 5 classes based on the exact time points when their expression levels peaked. These genes were labelled according to the cell cycle phases in which they peak (early G_1, late G_1, S, G_2, M.) As can been seen in figure 1 , some classes (e.g. late G_1 and S) display an overlap, while the S-class for example shows a large heterogeneity and can possibly be divided into two subclasses. It was further noted by Cho et al. that genes show frequently expression peaks in different classes. To achieve a more stringent assessment of the clustering performance, we introduce three statistical models of gene expression data for which we can control the structure and especially the number of clusters in the data. All these sets were based on the data set by Cho et al.

C. Model-based generation of gene expression data

The first model for gene expression data was based on a multi-variant Gaussian distribution. Each of the classes in this data set has the same mean vector, the same covariance matrix and the same number of genes as the corresponding class in the original data set by Cho et al. The second model-based data set was generated by random permutation of the expression values within the original class for each time point. This procedure yields classes that conserve the mean vector and the number of genes of the original classes, but not the covariance matrix. We also create a data set consisting of hyperspherical multi-variant Gaussian distributions. The covariance ma-

trices of the classes were constructed by multiplying the identity matrix with the average value of the diagonal elements in the original covariance matrix. Note that the first and the second model-based data set are based on hyperellipsoidal distribution. For the first data set, the axes of the hyperellipsoids are determined by the covariance of original classes, while for the second data set the axes parallel to the coordinate axes. Finally, baseline distributions were generated by random permutation of the gene expression values for each gene independently. Any correlation between genes in these data sets exist merely by chance.

III. Methods for Data Analysis

A. Preprocessing and normalization

The data was $\log_2$-transformed to achieve a symmetry between negative and positive fold changes and normalized to obtain a mean expression value of zero and standard deviation of expression values of one for each gene. This ensures that genes which share the same expression pattern have similar gene expression vectors.

B. Fuzzy c-means clustering

Clustering analysis seeks to achieve partitions of the data based on the similarity of the objects in the data. A partition divides the data into several clusters (or classes) and can be represented by a partition matrix U that contains the membership values μ_{ij} of each object i for each cluster j.

For hard clustering, which is based on classical set theory, clusters are mutually exclusive. This leads to the so called hard partitioning of the space. It is defined as

$$M_{hc} = \left\{ U_{ij} \in R^{c\times N} \;\middle|\; \begin{array}{l} \mu_{ij} \in \{0,1\} \;\forall\, i,j \\ \sum_{i=1}^{c} \mu_{ij} = 1 \;\forall\, j \\ 0 < \sum_{j=1}^{N} \mu_{ij} < N \;\forall\, i \end{array} \right\}$$

where c is the number of clusters and N the number of data objects.

Fuzzy clustering is based on the concept of fuzzy partitioning of the data space. In contrast to hard clustering, a data object can be member of several fuzzy clusters. This results in a fuzzy partitioned space that takes the form of

$$M_{fc} = \left\{ U_{ij} \in R^{c\times N} \;\middle|\; \begin{array}{l} \mu_{ij} \in [0,1] \;\forall\, i,j \\ \sum_{i=1}^{c} \mu_{ij} = 1 \;\forall\, j \\ 0 < \sum_{j=1}^{N} \mu_{ij} < N \;\forall\, i \end{array} \right\}$$

Note that the fuzzy partitioned space M_{fc} fully contains the hard partitioned space M_{hc} as a subspace.

Many different algorithms for cluster analysis aim to minimize an objective function. An important objective function for fuzzy clustering is the c-means functional J_m weighting the sum of squared errors within the clusters. In case of gene expression data, it can be written as

$$J_m(G,U,P) = \sum_{i=1}^{c}\sum_{j=1}^{N} (\mu_{ij})^m \|\mathbf{g_j} - \mathbf{p_i}\|_{\mathbf{A}}$$

where $\mathbf{g_i}$ is the expression vector of the gene i, U is the fuzzy partition matrix, $\mathbf{p_j}$ is the prototype (or cluster center) for cluster j, m the fuzzification parameter (with $m > 1$) and a distance norm $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{xAx}}$ takes the form of a squared inner product. Fuzzy c-means (FCM) clustering is the most common algorithm for solving this non-linear optimization problem. It is based on the first order conditions for a minimum of J_m:

$$u_{kl} = \frac{1}{\left(\frac{\|\mathbf{g_l}-\mathbf{p_k}\|_{\mathbf{A}}}{\sum_{i=1}^{c}\|\mathbf{g_l}-\mathbf{p_i}\|_{\mathbf{A}}}\right)^{\frac{2}{m-1}}} \quad \forall\, k,l \qquad (1)$$

$$p_k = \frac{\sum_{j=1}^{N}(\mu_{kj})^m \mathbf{g_j}}{\sum_{j=1}^{N}(\mu_{kj})^m} \quad \forall\, k \qquad (2)$$

A Picard iteration alternating between step (1) and (2) adjusts μ_{ij} and $\mathbf{p_j}$ until the change in J_m falls below a threshold ϵ or a maximal number of iterations t is reached. In this study, we chose $\epsilon = 0.001$ and $t = 100$. Note that p_j is the weighted mean of cluster j. The fuzzification parameter m controls the fuzziness of the partitioning *i.e.* the degree to which the membership of a gene is distributed among the clusters. For $m \to 1$, the fuzzy clustering turns into hard clustering of the data. The prototypes p_j are then simply the means of the clusters j. For $m \to \infty$, the partition approaches maximal fuzziness. A gene i is assigned to all clusters equally. We will analyze this behavior further and give recommendations for the choice of parameter m. Through setting the matrix A equal to the identity matrix, we selected the standard Euclidean norm as a distance. This choice has the advantage that no further parameters of matrix A have to be determined in our cluster analysis, however, we will see that it also leads to some drawbacks.

C. Cluster validity

Since we usually have little information about the data structure in advance, a crucial step in the cluster analysis is selection of the number of clusters. Finding the 'correct' number of clusters addresses the issue of cluster validity. This has turned out to be a rather difficult problem, as it depends on the definition of a cluster. Without prior information, a common method is the comparison of partitions resulting from different numbers of clusters. For assessing the validity of the partitions, several cluster validity functionals $f : U \to R$ have been introduced. These functionals should reach an optimum if the correct number of clusters is chosen. In this study we used the following validity functionals:

1. Partition coefficient F: Introduced by Bezdek [7], the

partition coefficient F is defined as

$$F(U) = \sum_{k,i=1}^{c,N} \mu_{ik}^2 / N$$

It is maximal if the partition is hard and reaches a minimum for $U = [1/c]$ when every object is equally assigned to every cluster.

2. Normalized partition coefficient $\tilde{F}$: It is well-known that the partition coefficient tends to decrease monotonically with increasing n. To reduce this tendency we define a normalized partition coefficient

$$\tilde{F} = F - F_0$$

where F_0 is the partition coefficient derived from the randomized data set.

3. Xie-Beni index S: Xie and Beni proposed an cluster validity index

$$S = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2 \|\mathbf{g_j} - \mathbf{p_i}\|^2}{n \underbrace{min}_{k \neq l} (\|\mathbf{p_k} - \mathbf{p_l}\|^2)}$$

that aims to quantify the ratio of the total variation within clusters and the separation of clusters. Choosing the correct number of clusters should minimize this index.

Using a fixed value for the fuzzification parameter m in the clustering validation process may give a wrong picture, since it is shown that the choice of m has a strong influence on the cluster validity functional presented here [8]. We therefore evaluate the cluster validity functionals by using a exhaustive grid-search method, varying both parameters c and m.

IV. Results

A. Determination of the FCM parameters

We performed FCM clustering for all data sets (i.e. the original data set selected by Cho et al. and the model-based data sets) in the parameter range of $2 \leq c \leq 10$ and $1.05 \leq m \leq 3.55$ and evaluated the performance of the cluster validity indices. The fuzzification parameter m turned out to be an important parameter for the cluster analysis. For the randomized data set, FCM clustering formed clusters only if m was chosen smaller than 1.15. Higher values of m led to uniform membership values in the partition matrix. This can be regarded as an advantage of FCM over hard clustering, which always forms clusters independently of the existence of any structure in the data. An appropriate choice for a lower threshold for m can therefore be set, if no cluster artifacts are formed in randomized data. An upper threshold for m is reached if FCM does not indicate any cluster in the original data. This threshold depends mainly on the compactness of the clusters. Compact clusters like the late G_1-cluster 'melt' for higher m than less compact clusters like G_2. Interestingly, cluster analysis of the model-based data sets showed

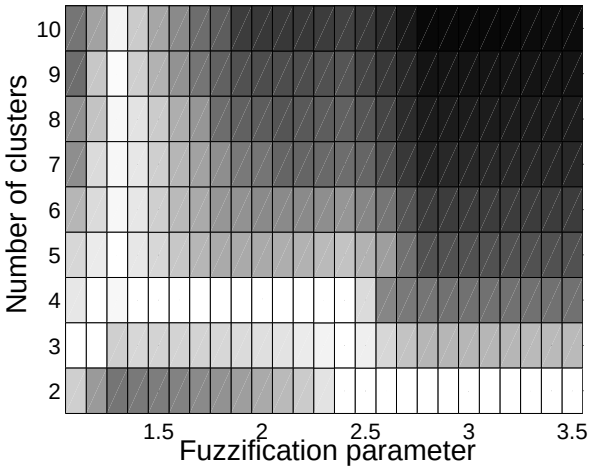

Fig. 2. Dependence of normalized partition coefficient $\tilde{F}$ on FCM parameter c and m for the original data set by Cho et al. Light gray values indicate high $\tilde{F}$; dark gray values indicate low $\tilde{F}$. For visualization purpose we scaled $\tilde{F}$, so that the maximum values for $\tilde{F}$ for any fixed m are the same.

that hyperspherical distributions are more stable for increasing m than hyperellipsoidal distributions. This may be expected since FCM clustering with Euclidean norm favors spherical clusters. Of all three compared validity indices only the normalized partition coefficient $\tilde{F}$ indicated five clusters in the data for a specific setting of m. The unmodified partition coefficient F reached its maximum for $c = 4$ for values of $m < 1.35$, while for larger m F showed a monotonic decrease in c. Xie-Beni's index S reached an minimum for $c = 4$ over the range of $1.05 \leq m \leq 2.55$. For larger values of m, $c = 2$ was indicated by S as the appropriate number of clusters. The main reason for S not pointing out five clusters, is strong dependence of the minimum distance of all cluster pairs. Analysis showed that the distance between the prototypes for the late G_1- and the S-cluster was much smaller compared to all other pair distances. Splitting these two clusters lead to a strong increase of S as only the minimum distance between cluster prototypes is used by S to evaluate the separation between clusters. Only the normalized partition coefficient $\tilde{F}$ indicated $c = 5$ in case of $m = 1.25$. Interestingly, the value of $\tilde{F}$ for $c = 5$ and $m = 1.25$ was also its global minimum. Higher values of m lead to smaller optimal number of clusters indicated by $\tilde{F}$. This behavior is shown in figure 2. For increasing m, $\tilde{F}$ reached its maximal value for smaller number of clusters.

B. Noise robustness

For the analysis of gene expression data, it is crucial that the applied clustering methods show a robust performance in the presence of a high level of noise. We compared the performance of fuzzy clustering with hard clustering by using two models to simulate noise. In the first model, uniformly distributed noise was added to the

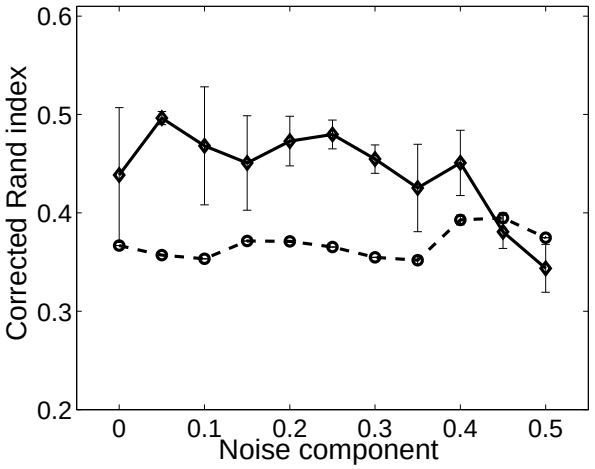

Fig. 3. Clustering performance dependent on percentage of added noise for the original data by Cho et al.: The continuous line displays the average and the standard deviation of corrected Rand indices achieved by k-means clustering, while the dash line corresponds to the average and standard deviation for FCM clustering. A noise component of 0.1 means that 10% of the each gene expression vectors consists of added uniform noise.

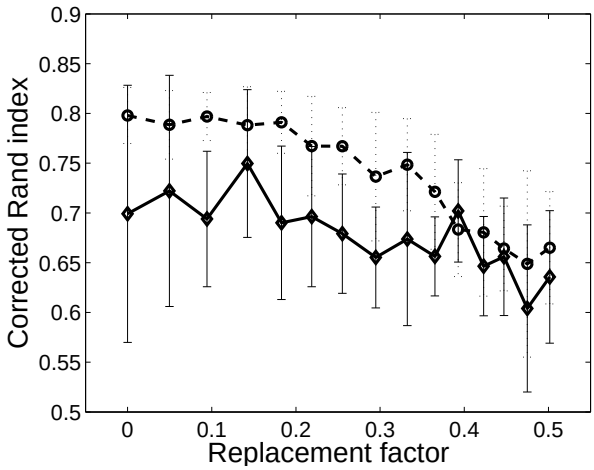

Fig. 4. Detection of clusters in a noisy environment based on the original data set by Cho et al.: Stability of partitioning was assessed by replacing genes by uniform noise. A replacement factor of 0.1 corresponds to a replacement of 10% of the genes by noise. The continuous line displays performance of hard clustering; the dash line displays performance for fuzzy clustering.

expression values of each gene. This intends to reflect the generic background noise in a microarray experiments. The second model contained a random selection of a chosen percentage of gene expression vectors which were replaced by uniform noise. This approach helps to assess the robustness of cluster analysis against data outliers and the sensitivity of finding patterns in a noisy environment. Both models are limited, since they assume a specific form of noise distribution. However, since no noise model has been established so far for microarray data, they may constitute a first step towards the assessment of the robustness of clustering microarray data. To assess the performance of the two clustering methods, we used the corrected Rand index [9]. This index compares two partitions and has a maximal value of 1 if the partitions are the same. For random partitioning the corrected Rand index yields 0. In short, it assesses the difference between two partitions.

In the first set of experiments, we calculated the corrected Rand indices comparing the classes of the original data set with the partitions achieved by the clustering methods when noise is added to the data. Uniform random noise was added stepwise to every gene expression vector up to a level of 50%. For every noise level, 10 runs with random initiation were performed for both clustering methods. The fuzzification parameter m was chosen to be 1.1, since a better performance of FCM clustering was observed for small values of m in the case of a large noise component. The number of clusters was set to 5 for both methods. For calculation of the Rand index, it is necessary to convert the fuzzy partition to a hard partition. This was achieved by assigning genes to the cluster with the largest membership value. For the original data set by Cho et al., k-means clustering outperformed fuzzy c-means clustering for a percentage of added uniform noise up to 40% (see figure 3). Two points, however, are important to note. First, k-means clustering produced generally different partitions in repeated runs and thus showing a large variation in the Rand index. This can be regarded as a drawback of k-means clustering, as usually no external criterion exists to select the specific partition from different runs. FCM yielded very stable results as it converges to very similar partitions in different runs even for a large noise component. Second, the performance of k-means is more affected by increased noise and decreases strongly for high levels of noise. FCM achieves similar Rand index for different noise levels and outperforms k-means clustering for noise levels over 40%. To explain why FCM clustering performs worse than k-means in case of low levels of noise we analyzed the clustering of the model-based data sets. The performances of both clustering methods applied to the model-based data with conserved covariance structures were similar to the results for the original data. Both methods, however, improved strongly if model-based data with diagonal covariance was used. K-means still performed better over a wide range of noise levels, while FCM resulted in a more stable partitioning of the data. This is contrasted by the results for the modelled data based hyperspherical normal distributions. FCM showed a better and more stable performance than k-means. This indicates that FCM with Euclidean distance is rather sensitive to the shape of clusters.

In the next set of experiments we compared the partition achieved for the data sets without added noise to the partitions for data with added noise. This may give

0-7803-7280-8/02/$10.00 ©2002 IEEE

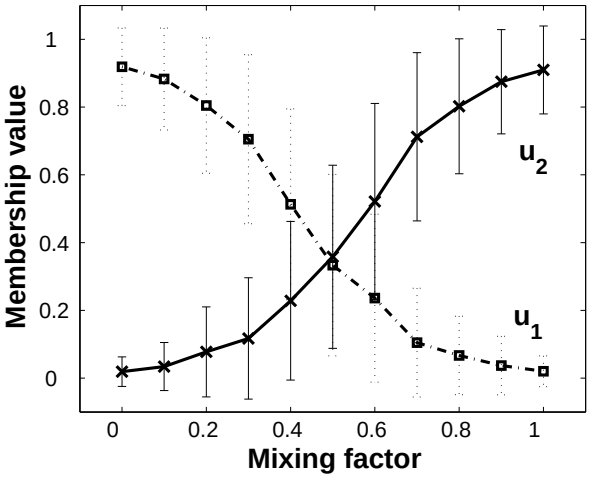

Fig. 5. Average membership function value of model-based generated gene vectors in dependence of the mixture of two clusters.

a fairer comparison of the two clustering methods since it assesses the stability based on an internal criterion. We replaced stepwise a fraction of randomly selected genes by uniform distributed noise (see figure 4). For all data sets FCM showed more stability. The difference in performance becomes marginal if more than a third of the genes are replaced by noise.

C. Multiple parents cluster

In the final part of this study, we assessed the capability of FCM clustering to detect genes that may be members of two clusters simultaneously. Since the study of genome-wide expression patterns is still in its infancy and to date no gene expression data sets exists for which the underlying regulatory network structure is fully known, we modeled genes belonging to two clusters based on multivariate normal distributions. The centers and the covariance matrices of these modeled distributions were the weighted sums of the prototypes and covariance matrices of existing clusters that were randomly selected. The contribution of each existing cluster to the simulated gene expression vectors were controlled by a weighting factor. In this experiment, we generated 40 genes for different settings of the weighting factor and recorded the results of FCM clustering for 10 independent runs for each setting. The membership values of the modeled gene expression vectors for the clusters they were derived from were recorded and are displayed in figure 5. These membership values correspond well with the mixing factor which determines the contribution of each mother cluster. The analysis of membership values may therefore have the potential to reveal multiple regulations of genes in future studies.

V. Conclusions

The field of bioinformatics has recently been attracting more and more attention. A vast variety of different algorithms has been introduced to this new field. However, fuzzy methods have not been able to establish themselves as tools for data analysis in bioinformatics. We have shown in this work that fuzzy concepts can be regarded as a promising and a powerful representation of complex biological data structures. We analyzed here mainly the parameter selection and the robustness of fuzzy clustering against background noise and outliers. This analysis is a first approach towards exploiting the robustness of fuzzy clustering to gain new insights in gene expression data. Although several cluster methods have recently been applied to gene expression data, the assessment of the robustness of cluster methods has been neglected, mainly as researchers have concentrated so far on a few dominated patterns in gene expression data. Robust cluster analysis, however, is of crucial importance to discover potential important subtle patterns that are difficult to distinguish from background noise. Many topics of research remain to be resolved. To name just a few: adjustment of clustering parameters to the data, choosing an appropriate similarity measure, finding suitable distance norms, statistical significance of clusters, biological validation of fuzzy clusters.

VI. Acknowledgements

We have applied the fuzzy clustering approach to several complete gene expression data sets. The discussion of the underlying biology, however, is beyond the scope of this study that focusses on the technical issues of fuzzy clustering applied to microarray data. The main findings of the clustering analysis of the complete data sets will therefore be published elsewhere [10].

References

[1] J.L.DeRisi, V.R.Iyer and P.O.Brown," Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, Vol.275, pp. 680-686,1997

[2] S. Chu, J.L.DeRisi, M.B.Eisen, J. Mulholland, D.Botstein, P.O.Brown, B.Futcher, " The transcriptional program of sporulation in budding yeast", *Science*, pp.699-705, 1998

[3] S.Tavazoie, J.D.Hughes, M.J.Campbell, R.J.Cho, G.M.Church," Systematic determination of genetic network architecture", *Nature genetics*, vol.22, pp.281-285, 1999

[4] P.Törönen, M.Kolehmainen, G.Wong, E.Castrén," Analysis of gene expression data using self-organizing maps", *FEBS Letters*, Vol.451, pp. 142-146,1999

[5] M.B.Eisen, P.T. Spellman, P.O.Brown, D.Botstein," Cluster analysis and display of genome-wide expression patterns", *PNAS*, vol.95, pp. 14863-14868, 1998

[6] R.J.Cho, M.J.Campbell, E.A.Winzeler, L.Steinmetz, A.Conway, L.Wodicka, T.G.Wolfsberg, A.E. Gabrielian, D. Landsman, D.J.Lockhart, R.W.Davis," A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol.2, pp.65-73, 1998

[7] J.C.Bezdek, Pattern recognition with fuzzy objective function algorithms, Plenum, 1981The transcriptional program of sporulation in budding yeast. Science. 1998 Oct 23;282(5389):699-705.

[8] N.R.Pal, J.C.Bezdek," On cluster validity for the fuzzy c-means model", *IEEE Trans. on fuzzy systems*, pp.370-379, 1995

[9] L.Hubert, P.Arabie,"Comparing partitions", *J.Classification*, pp.193-218, 1985

[10] M.E.Futschik, N.K.Kasabov," Robust soft clustering of gene expression data", submitted to *Bioinformatics*