International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:5, 2007

Evaluation of Clustering based on Preprocessing in Gene Expression Data

Seo Young Kim, and Toshimitsu Hamasaki

***Abstract*—**Microarrays have become the effective, broadly used tools in biological and medical research to address a wide range of problems, including classification of disease subtypes and tumors. Many statistical methods are available for analyzing and systematizing these complex data into meaningful information, and one of the main goals in analyzing gene expression data is the detection of samples or genes with similar expression patterns. In this paper, we express and compare the performance of several clustering methods based on data preprocessing including strategies of normalization or noise clearness. We also evaluate each of these clustering methods with validation measures for both simulated data and real gene expression data. Consequently, clustering methods which are common used in microarray data analysis are affected by normalization and degree of noise and clearness for datasets.

***Keywords*—**Gene expression, Clustering, Data preprocessing.

I. Introduction

MICROARRAYS allow the high-throughput analysis of gene information, and thus have led to revolutionary changes in bioinformatics research. Microarray technology is being applied to biological and medical investigations of the reliable and precise classification of tumors, which is essential to the successful cancer treatment. A critical aspect of such analyses of microarray data is the identification of classes of genes with similar function. An important statistical problem in tumor classification is identifying new tumor classes based on gene expression profiles. To obtain the goal, the clustering is the basic tool that has been applied to microarray data. Due to the vast number of genes involved in microarray experiments and the complexity of biological processes, an effective clustering algorithm for grouping samples is essential for such studies like tumor classification, function annotation, and other biomedical applications [1].

Clustering faces two main problems. One is how to correctly determine the number of clusters, and the other is how to assign samples to those clusters. Since a clustering analysis relies heavily on limited biological and medical information like tumor classification, the results therefrom are not only sensitive to noise but also prone to over-fitting, which is a major concern in the clustering analysis of microarray data. Many studies have analyzed microarray data with the aim of identifying sample classes using methods like hierarchical clustering (HC), *K*-means (KM) [2], ensemble clustering [3],[4], partitioning around medoids (PAM) [5], fuzzy C-means (FCM) [6]. These clustering methods can be broadly divided into two groups: (1) partitioning methods like the FCM, PAM, and KM, which seek to optimally dissect objects into a fixed number of clusters even if the expression profile of each sample has a number of similar cluster patterns, and (2) hierarchical methods, which produces a nested sequence of clusters [7].

In medical cancer diagnoses based on microarray data, the definition of tumor classes would be based on clustering results[8]. Inaccurate cluster assignments could lead to misdiagnoses and poor treatment protocols. Therefore, statistical clustering is very important to identifying new tumor classes. The results of clustering depend on various characteristics of the data, including the microarray experimental conditions, the variations between data points, and the degree of noise. Data preprocessing is therefore an essential procedure for handling original microarray data, but this has often not been used in cluster analyses thereof. Although many studies have compared the performances of various clustering methods [9],[10], such comparisons could be invalid without the inclusion of data preprocessing. In particular, many studies mainly used that all data are normalized such that every sample has a mean expression value of 0 and a standard deviation (SD) of 1 across genes [8],[9],[11]. However, this is simply no more than location and spread transformation of data though it was commonly used in microarray data analysis, in what ways, this method also could reflect no transformation. It can not be the most transformation when the data include the complicated variability from two directions of genes and samples like microarray data [12]. Thus, this study especially shows how the clustering result is not good when the simple mean and SD normalization is applied.

In this paper, we compare the performances of the HC, FCM, PAM, and KM whilst considering data preprocessing by normalization, effective gene selection among thousands of

Manuscript received December 14, 2007. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2005-214-C00186).

S. Y. Kim is with Statistical Research Institute, Korea National Statistical Office, Daejeon, 302-120 Korea (corresponding author to provide phone: +82-42-717-0215; fax: +82-42-717-0251; e-mail: sykim2217@ nso.go.kr).

Hamasaki. T is Biomedical Statistics, Osaka University Graduate School of Medicine, 2-2 Yamada-oka, Suita, Osaka 565-0871, Japan (e-mail: hamaskt@medstat.med.osaka-u.ac.jp).

genes, and noise treatment. The performances of the four clustering methods are compared using simulated and real microarray datasets, with the results evaluated based on the silhouette (*sil*) [5], and the adjusted Rand index [13].

II. Materials and Methods

A. Datasets

We used five simulated datasets such as Sim 1 to Sim 5, and three real gene expression datasets such as leukemia [11], melanoma [14], and lung cancer [15] to inspect the performance and features of clustering methods based on the data preprocessing.

Sim 1. Three clusters in two dimensions: 25, 25, and 50 objects are generated from bivariate normal distribution in each of the three clusters with means (0, 0), (0, 5), and (5, − 3), respectively, and 2**I** covariance matrix, where the matrix **I** is an identity matrix.

Sim 2. Four overlapping clusters in 10 dimensions: each cluster is chosen to have 50 objects from normal distribution with an appropriate mean vector and an identity covariance matrix. The cluster means are randomly chosen from a bivariate normal distribution $N_2(\mathbf{0}, 2.5\mathbf{I})$. Each simulated where the Euclidean distance between the two closest objects belonging to different clusters is less than 1 discarded.

Sim 3. Two elongated clusters in three dimensions: cluster 1 is generated as follows. Set $x_1 = x_2 = x_3 = t$ with t taking on 100 equally spaced values from -0.5 to 0.5, and then let Gaussian noises with standard deviation 0.1 be added to each variable. Cluster 2 is generated in the same way except that the value 10 is added to each variable. These result in elongated clusters, stretching out along the main diagonal of a three dimensional cube.

Sim 4. Three overlapping clusters in 13 dimensions, 10 noise variables: the first three variables in each of three clusters have a multivariate normal distribution with mean vectors (0,0,0), (2, -2, -2), and (-2, 2, -2), respectively, and with covariance matrix Σ, where $\sigma_{ii} = 1$, $1 \le i \le 3$, and $\sigma_{ij} = 0.5$, $1 \le i \ne j \le 3$. The remaining 10 noise variables are generated independently from the $N_{10}(\mathbf{0}, \mathbf{I})$ distribution. Each cluster contains 50 objects.

Sim 5. Two overlapping clusters in 10 dimensions, 9 noise variables. Each cluster contains 50 objects. The first variables in each cluster were generated from normal distribution with mean 0 and 2.5, respectively, and with variance 1. The remaining nine noise variables are generated from the $N_9(\mathbf{0}, \mathbf{I})$ distribution.

Note that above configurations in generating the datasets were considered in earlier papers [8],[16]. The configurations are summarized in Table I.

TABLE I
Description of Five Simulated Datasets

Data	true clusters	Number of dimensions	Number of objects in each cluster	Degree of overlap among clusters
Sim1	3	2	25, 25, 50	None
Sim2	4	10	50, 50, 50, 50	Strong
Sim3	2	3	100, 100	None
Sim4	3	13	50, 50, 50	Strong
Sim5	2	10	50, 50	Weak

Leukemia. Leukemia dataset consists of 38 learning samples on the Affymetrix high density oligonucleotide chips containing 7129 human genes [11]. The goal of this experiment is to identify genes that are differentially expressed in 27 acute lymphoblastic leukemia (ALL) patients and 11 acute myeloid leukemia (AML) patients.

Melanoma. The melanoma dataset consists of two types of 31 cutaneous melanomas and 7 controls [14]. Gene expression levels were measured using cDNA microarrays containing 8150 human genes. Of the 8150 genes, 3613 genes were identified as well measured. This experiment data had many Cy5/Cy3 expression ratios above 10000 and also had many below 0.02. Therefore, the data filtering method, which excluded the genes with expression ratio greater than 50 and less than 0.02, was applied for this dataset[17]. These ratios were transformed to a base 2 logarithmic scale.

Lung cancer. The lung cancer dataset comes from a study gene expression in five types lung carcinoma: 139 lung adenocarcinomas, 21 squamous cell lung carcinomas, 20 pulmonary carcinoids, 6 small-cell lung carcinomas cases, and 17 normal lung specimens. Gene expression levels were measured using Affymetrix high density oligonucleotide arrays containing 12000 human genes [15].

B. Preprocessing Strategy

Normalization

It is meaningful to remove from microarray data variations due to non-biological factors [18]. This process, known as normalization, is important for obtaining reliable data for subsequent analysis. One of the most commonly utilized normalization approaches is the mean and SD normalization, whereby all data are normalized such that every object has a mean expression value of 0 and a SD of 1 across genes. And, one of often used nonlinear correction methods is locally weighted scatter plot smoothing (LOWESS), which was first applied to microarray data by Yang *et al*. [12]. The main idea of LOWESS is to utilize a locally weighted polynomial regression of the intensity scatter plot to obtain the calibration factor. Compared to other methods, the LOWESS method is known to be robust across a wider range of types of datasets.

Preliminary Gene Selection

Thousands of gene expression levels were monitored in each of the three microarray datasets. But, the large numbers of genes exhibit nearly constant expression levels, as measured by the variance of the expression levels across arrays. These genes did not seem to be useful for classification purposes, and thus we used genes with high-variance of expression levels across objects in the clustering process. First three plots of Fig. 1 show for each dataset the individual gene variance divided by the maximum variance over all genes [8]. In these plots, all variance curves show a clear drop-off which gradually flattens. All plots indicate similar patterns for three microarray datasets. In this paper, we selected 100 genes with high variance from Leukemia and Melanoma datasets, and then we selected 200 genes for Lung cancer dataset as it contains more classes [8]. The last box plot of variance of Fig. 1 indicates the dispersion of each of three datasets. Of these, melanoma dataset has more variance for expression levels relative to other two microarray datasets.

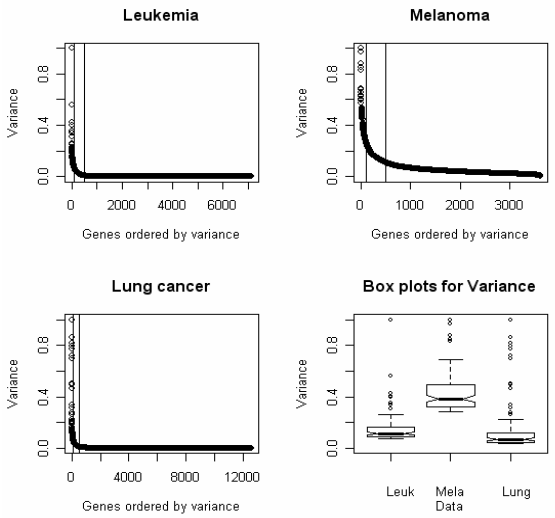

Fig. 1 Plots of the variance of the expression levels of each gene across samples. The variances are scaled by the maximum variance over all genes, and the genes are ordered by variance in descending order. The vertical lines indicate 100 and 500 genes. The last plot is the box plot of the variances for the three microarray datasets

C. Clustering Algorithm

Fuzzy C-Means Clustering

The FCM can be represented as follows [6]:

$$minimize\ J_{fcm}(W,V)=\sum_{i=1}^{n}\sum_{k=1}^{c}(w_{ik})^{m}\,\|x_i-v_k\|^2,$$

where, $J_{fcm}(W,V)$ represents the objective function defining the quality of the result obtained for prototypes V and membership W, and m is the degree of fuzziness in the clustering. The membership degrees w_{ik} are defined such that $0\le w_{ik}\le 1$, under the constraint of $\sum_{k=1}^{c} w_{ik}=1$ for $i=1,\ldots,n$. $V=(\boldsymbol{v}_k)$ is the cluster center or prototype, and $\|\boldsymbol{x}_i-\boldsymbol{v}_k\|^2$ is the squared Euclidean distance between object i and the prototype of cluster k. Here, $\boldsymbol{x}_i$ indicates i*-th* object vector. In case of the FCM clustering, when they are applied to microarray data, it is very important to choose appropriate values for the fuzziness parameter due to its effect on the minimization criterion for the objective function. Dembele and Kastner showed that the common value of 2 used for the fuzziness parameter is not appropriate for microarray data [19]. For the FCM, we firstly determined the optimal fuzziness parameter [19], and then used it when implementing the FCM algorithm. In general, the FCM is known to provide poor results compared to other crisp clustering methods when preprocessing is not considered [9]. However, here we show that the method produces accurate and clear clustering results. Although we do not mention for the computational procedure about fuzziness parameter, we can refer to [19] about it for the detail.

Hierarchical Clustering

The agglomerative HC is one of popular methods for clustering gene expression. The clustering is based on a pairwise distance matrix between objects. The nested sequence of clusters produced by the HC makes them appealing when different levels of detail are of interest because small clusters are nested inside larger ones. The result of the algorithm is a dendrogram, which shows how the clusters are related. A clustering of the samples into disjoint groups is obtained by cutting the dendrogram at some level. In microarray data analysis, the goal of clustering may focus on both small groups of similar samples and a few large clusters. The former occur when small samples have special meanings, and the latter occur when larger groups exist, such as samples from several sources, or from different experiments [7].

PAM Clustering

The PAM is a partitioning algorithm and can be regarded as a generalization of *k*-means clustering to arbitrary dissimilarity matrices [5]. The PAM is based on the search for *k* representative medoids, among the objects to be clustered. After finding this *k* medoids, *k* clusters are built by allocating each object to the nearest medoid. This goal is to minimize the sum of the dissimilarities of the objects to their cloest medoid. The algorithm consists of two steps. The *k* initial sets of medoids are firstly sequentially selected, and then swap points so that the objective function is minimized iteratively by replacing one medoid with another entry and this step is repeated until convergence. The PAM is known to be more robust and computationally efficient than the KM [8].

K -means Clustering

The most popular partitioning method, the KM clustering partitions data into *k* clusters such that objects in the same cluster are more similar to each other [2], *i.e.*, the clusters are

internally similar, but externally dissimilar. The goal is to divide the objects into k clusters such that some metric relative to the centroids of the clusters is minimized. Two steps are available to search for the optimum set of clusters. The algorithm first assigns each object to a cluster that has the closest centroid, and then sets initial positions for the cluster centroids, that is, when all objects have been assigned, recalculate the positions of the k centroids. This procedure is continued until the optimum assignment of objects to clusters is found [20]. In this process, the KM method should have little difficulty with missing data because mean updates and distance computations can be performed with some missing values.

D. Evaluation Criteria

Sil index: The method is to select the number of clusters which gives the largest average silhouette width, $ave\ sil_j = \sum_{i=1}^{n_j} s(i)/n_j$, where n_j is the number of objects in the j^{th} cluster. The silhouette width for the i^{th} object in the j^{th} cluster is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

Here, $a(i)$ is the average distance between the i^{th} object and all of the objects clustered in the j^{th} cluster, and b(i) is the smallest average distance between the i^{th} object and all of the objects clustered in cluster l ($1 \le j, l \le k, j \ne l$) [3].

Adjusted Rand index: This method computes the extent of agreement between two partitions. Given the set $D=\{o_1, o_2, \ldots, o_n\}$, suppose $U=\{u_1, u_2, \ldots, u_R\}$ and $V=\{v_1, v_2, \ldots, v_C\}$ represent two different partitions of the objects in D. Here, for $1 \le i \ne i' \le R$ and $1 \le j \ne j' \le C$, $\bigcup_{i=1}^{R} u_i = \bigcup_{j=1}^{C} v_j = D$ and n_{ij} is the number of objects that are in both classes u_i and v_j, and n_i and n_j are the number of objects in classes u_i and v_j, respectively. The adjusted Rand index is as follows [13]:

$$Rand = \frac{\sum_{ij} {}_{n_{ij}}C_2 - \left[\sum_i {}_{n_{i.}}C_2 \sum_j {}_{n_{.j}}C_2\right] / {}_nC_2}{(1/2)\left[\sum_i {}_{n_{i.}}C_2 + \sum_j {}_{n_{.j}}C_2\right] - \left[\sum_i {}_{n_{i.}}C_2 \sum_j {}_{n_{.j}}C_2\right] / {}_nC_2}.$$

For good clustering, we expect these values to be high [3].

III. Analysis Results

The clustering methods described in ealier section were applied to the three real microarray datasets and five simulated datasets. To compare the four clustering methods, we selected the more variable genes across arrays from the microarray dataset [8], and then applied both normalization methods to all the datasets. The values of the fuzziness parameters used for the FCM clustering of all datasets are given in Table II. The degree of fuzziness differed markedly between the datasets, being low for datasets with clear clustering and small variations such as Sim1, Sim3, and Lung cancer, and being lower for mean and SD normalization than for the LOWESS normalization. Therefore, no general assumption can be made for values of fuzziness that might give good FCM clustering results with a microarray dataset. This result leads us to the hypothesis that applying data preprocessing prior to clustering will greatly affect the results of clustering.

TABLE II
FUZZINESS PARAMETERS USED FOR THE FCM CLUSTERING.

Dataset	true clusters	LOWESS	Mean and SD
Sim 1	**3**	1.35	1.27
Sim 2	**4**	1.57	1.28
Sim 3	**2**	1.30	1.25
Sim 4	**3**	1.54	1.20
Sim 5	**2**	1.76	1.24
Leukemia	**3**	1.22	1.12
Melanoma	**2**	1.23	1.11
Lung cancer	**5**	1.15	1.13

Two aspects of the performance of each clustering method were evaluated using both the *sil* and the adjusted Rand indices: (1) finding the correct number of clusters and (2) correctly assigning each object to the resulting clusters. Furthermore, to test the ability of each clustering method to handle noisy data, we performed additional clustering calculations on microarray datasets to which noise had been added, with subsequent the LOWESS normalization. The noisy datasets were generated by adding a small Gaussian random variable with 0 mean and 1.5 SD to the expression levels in the original microarray dataset.

A. Results for the Simulated Datasets

We performed an extensive simulated study to evaluate each of the FCM, HC, PAM, and KM clustering methods.

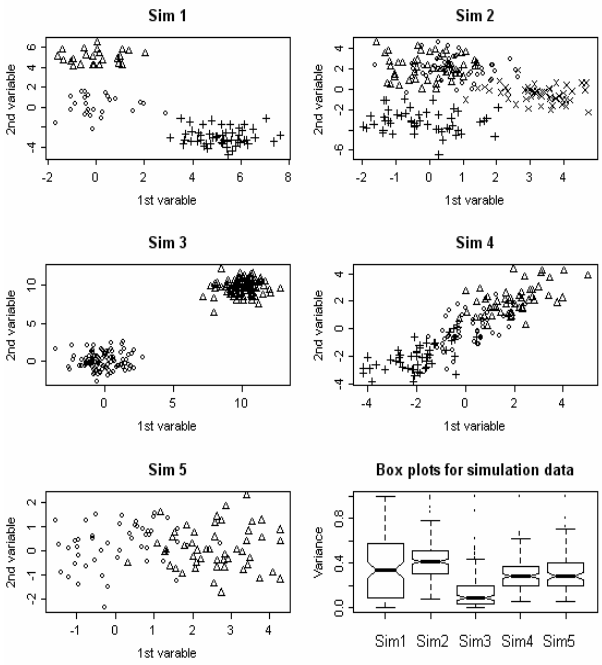

Fig. 2 Cluster distributions for the five simulated datasets, and a box plot showing the variation in the data points between and within clusters for each dataset

Above Fig. 2 depicts the shape of the cluster for each of the simulated datasets. Sim 1 and Sim 3 clearly exhibit three and two clusters, respectively, with very small variation in each cluster. As shown in box plot of Fig. 2, Sim 1 exhibits a large variation between data points, this may result in Sim 1 has a larger variation between clusters than within clusters. Sim 2 has heavily overlapped clusters with four clusters and unclear boundary between clusters. Sim 4 also has heavily overlapped clusters, with only small deviations in the data points from a linear trend. Thus these Sim 1 and Sim 2 are not expected to be accurate estimations of the clustering. Sim 5 has two clusters, and they are slightly unclear and more variable within each cluster compare to the other datasets.

Table III lists the results of estimating the number of clusters using *sil* for the five simulated datasets. We first describe the results from applying LOWESS normalization. As indicated in Table III, all the clustering methods correctly identified the number of clusters for Sim 1 and Sim 3 (with clear clusters) and Sim 5 (with large variations between data points). However, for Sim 2 (with heavily overlapping among the clusters), only the FCM and KM were able to correctly classify the clusters. All the clustering methods failed to find the correct number of clusters for Sim 4, presumably due to the degree of overlap and linearity between the clusters. On the other hand, when applying the mean and SD normalization, most of the clustering methods failed to find the correct number of clusters for all datasets except Sim 3 and Sim 5. In particular, for mean and SD normalization, all the clustering methods estimated that there were two clusters in Sim 1, whereas they correctly estimated the presence of three clusters when LOWESS normalization was applied. These results indicate that the clustering methods are greatly affected by the degree of overlap. Moreover, the FCM and KM clusterings are stable irrespective of the shape of the data, and LOWESS normalization could be refined to cope with the presence of noise in the dataset.

TABLE III
ESTIMATING THE NUMBER OF CLUSTERS IN SIMULATED DATA USING THE *SIL* PROCEDURE

		LOWESS				**Mean and SD**			
data	ture	FCM	HC	PAM	KM	FCM	HC	PAM	KM
Sim 1	**3**	**3**	**3**	**3**	**3**	2	2	2	2
Sim 2	**4**	**4**	3	3	**4**	2	2	2	2
Sim 3	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Sim 4	**3**	2	2	2	2	2	2	2	2
Sim 5	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**

Table IV lists the adjusted Rand indices for simulated datasets. In aspect of the normalization method, all the clustering methods showed better performance for LOWESS normalization than for mean and SD normalization for all datasets. In LOWESS normalized datasets, the FCM, and KM clusterings outperformed the other clustering methods, and in mean and SD normalized datasets the FCM and PAM clusterings perform better than other two clustering. These results indicate that the clustering method is sensitive to data normalization and the degree of variation or in overlapping between clusters.

TABLE IV
ADJUSTED RAND INDICES FOR THE KNOWN CLUSTERS IN THE SIMULATED DATASETS FROM EACH CLUSTERING METHOD

		LOWESS				**Mean and SD**			
Data	true	FCM	HC	PAM	KM	FCM	HC	PAM	KM
Sim 1	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.42
Sim 2	4	0.93	0.88	0.81	0.62	0.56	0.45	0.54	0.61
Sim 3	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sim 4	3	0.60	0.31	0.34	0.53	0.33	0.21	0.30	0.26
Sim 5	2	0.57	0.57	0.49	0.57	0.37	0.19	0.38	0.20

B. Results for the Real Microarray Datasets

We next evaluated the four clustering methods using three gene expression datasets whilst applying only LOWESS normalization due to the results of simulated datasets. To check the ability of clustering methods dealing with nose data, we compared the performances of the clustering methods on noisy datasets. As mentioned in earlier, the noisy datasets were generated by adding a small Gaussian random variable with 0 mean and 1.5 SD to the expression levels in the original microarray dataset. Table V lists the estimations of the numbers of clusters using *sil* for original and noise microarray datasets, and Table VI lists the corresponding adjusted Rand indices. Tables V and VI show that for the original leukemia and melanoma datasets with small numbers of clusters, all the clustering methods correctly identified the number of clusters. Moreover, the FCM and PAM clusterings found the correct numbers of clusters in the noisy datasets. However, none of the clustering methods found the correct number of clusters for the lung cancer dataset, which contained five clusters. This may be due to the dataset containing complicated and overlapping clusters.

TABLE V
ESTIMATING THE NUMBER OF CLUSTERS IN ORIGINAL AND NOISY MICROARRAY DATASETS USING THE *SIL* PROCEDURE

		Original data				**Noisy data**			
Data	true	FCM	HC	PAM	KM	FCM	HC	PAM	KM
Leukemia	3	**3**	**2**	**3**	**3**	**3**	2	**3**	2
Melanoma	2	**2**	**2**	**2**	**2**	**2**	3	**2**	**2**
Lung cancer	4	3	2	2	2	2	2	2	2

The FCM and KM clustering methods exhibited higher accuracy for the original leukemia and lung cancer datasets. However, the FCM and PAM outperformed the other clustering methods for the two noisy datasets, as indicated by the adjusted Rand indices given in Table VI. Like Sim 5, which had a small number of clusters but large variations, the melanoma dataset contained more variable data points than did the other datasets (Fig. 1). Tables V and VI also indicate that for the melanoma dataset, all the clustering methods correctly identified the number of clusters, but the values of adjusted Rand indices

were very low. To confirm the effects of noise for data, Table VII compares the agreement of clustering between the original and noisy datasets using the adjusted Rand index for each clustering method. In the Table VII, the FCM and PAM clusterings exhibited good agreement of clustering performances for noisy datasets, with the KM and HC being slightly sensitive to the noisy data for all real datasets.

TABLE VI
COMPARISON RESULTS BY ADJUSTED RAND INDEX CORRESPONDING TO TRUE CLUSTERS FROM AN ORIGINAL AND NOISY MICROARRAY DATASETS

		Original data				Noisy data			
Datasets	true	FCM	HC	PAM	KM	FCM	HC	PAM	KM
Leukemia	**3**	0.69	0.41	0.45	0.69	0.60	0.39	0.44	0.22
Melanoma	**2**	0.31	0.31	0.33	0.31	0.20	0.27	0.26	0.16
Lung cancer	**4**	0.58	0.38	0.53	0.55	0.58	0.37	0.53	0.50

TABLE VII
COMPARISON RESULTS BY ADJUSTED RAND INDEX BETWEEN THE CLUSTERING FOR BOTH ORIGINAL AND NOISY DATASETS FOR EACH METHOD

Datasets	true	FCM	HC	PAM	KM
Leukemia	**3**	1.00	1.00	1.00	0.80
Melanoma	**2**	0.60	0.23	0.52	0.57
Lung cancer	**4**	0.96	0.87	0.94	0.79

IV. CONCLUDING REMARKS

We have compared the performance of various clustering methods using both simulated and real microarray datasets. In this paper, we considered various properties of datasets and data preprocessing procedures, and have confirmed the importance of preprocessing microarray data prior to performing core cluster analysis. All clustering methods were affected by data normalization and the data characteristics, such as the overlapping between clusters and the presence of noise. In particular, we identified that the mean and SD normalization, which is commonly applied to microarray data, was not appropriate than the LOWESS normalization in microarray data. Many other normalization methods that could be also applied [12],[18],[21] as the alternative of mean and SD normalization. We impress that the first step in explanatory clustering of microarray data should be to normalize the data so as to remove systematic variations due to non-biological factors.

Also, our comparative investigations revealed that the FCM and PAM clusterings were generally more accurate and consistent, in terms of finding the correct number of clusters and of assigning almost all objects to the correct clusters. Moreover, their cluster assignments tended to be more accurate when we used noisy data, as assessed by the adjusted Rand indices. The KM and HC clusterings tended to be not robust to both noisy data and data with overlapping clusters. The FCM showed particularly stable results for datasets with noise and overlapping clusters when we used the optimal fuzziness parameter. The crisp clustering methods like the PAM, HC, and KM forcibly assigns all genes to clusters, even those for which the variations in expression do not fit into any global pattern,[19] while in the FCM clustering each gene can belong to more than one cluster, with a graded association with each cluster, only one of which may be biologically significant [19]. If one focus on finding genes showing coherent behavior within clusters the FCM clustering could be useful.

REFERENCES

[1] J. Quanckenbush, "Computational analysis of microarray data," *Nat.Genet.* vol. 2, 2001, pp. 418-427.
[2] J. A. Hartigan, M. A. Wang, "A k-means clustering algorithm," *Appl.Stat.* vol.28, 1979, pp. 100-108.
[3] S. Y. Kim, J. W. Lee, "Ensemble clustering method based on the resampling similarity measure for gene expression data," Statistical methods in medical research, vol. 16, 2007, pp. 539-564.
[4] A. Weingessel, E. Dimitriadou, K. Hornik, "An ensemble method for clustering," *DSC Working papers,* 2003. See also http://www.ci.tuwien.ac.at/Conferences/ DSC-2003.
[5] L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, New York, 1990.
[6] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.
[7] T. Speed, *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall, New York, 2003.
[8] S. Dudoit, J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset,".*Genome Biology,* vol.3, 2002, research0036.1-0036.21.
[9] S. Datta, S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics* vol.19, 2003, pp. 459-466.
[10] Y. Luan, H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics* vol.19, 2003, pp. 474-482.
[11] 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, E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science* vol. 286, 1999, pp. 531-537.
[12] Y. H. Yang, S. Dudoit, P. Luu, T. P. Speed: Normaliztion *for cDNA microarray data*, eds. M. Bittner, Y. Chen, A. Dorsel, E. Dougherty, *Microarrays: Optical Technologies and Informatics SPIE,* 2001.
[13] K. Y. Yeung, W. L. Ruzzo, "An empirical study on principal component analysis for clustering gene expression data," *Technical Report 2000 UW-CSE-00-11-01*, Department of Computer Science and Engineering, University of Washington, 2001.
[14] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature* vol.406, 2002, pp. 536-540.
[15] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas sub-classes," *Proc.Natl. Acad.Sci*. vol. 98**,** 2001, pp. 13790-13795.
[16] R. Tibshirani, G. Walther, T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *Technical Report, Department of Biostatistics, Stanford University*, 2000.
[17] R. G. Darlene, G. Debashis, M. C. Erin, "Statistical issues in the clustering of gene expression data," *Statistica Sinica* vol.12, 2002, pp. 219-240.
[18] Y. Zhao, M. C. Li, R. Simon, "An adaptive method for cDNA microarray normalization," *BMC Bioinformatics* vol. 6; 28, 2005.
[19] D. Dembele, P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics* vol. 19, 2003, pp. 973-780.
[20] V. Guralnik, G. Karypis, "A scalable algorithm for clustering protein sequences," *Workshop on Data Mining in Bioinformatics*, *Proceedings of the U.S.A*., 2001, pp. 73-80.
[21] J. A. Berger, S. Hautaniemi, A. K. Jarvinen, H. Edgren, S. K. Mitra, J. Astola, "Optimized LOWESS normalization parameter selection for DNA microarray data," *BMC Bioinformatics* vol. 5, 2004, pp. 194.