World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:1, No:4, 2007

Fuzzy Types Clustering for Microarray Data

Seo Young Kim, Tai Myong Choi

Abstract**—**The main goal of microarray experiments is to quantify the expression of every object on a slide as precisely as possible, with a further goal of clustering the objects. Recently, many studies have discussed clustering issues involving similar patterns of gene expression. This paper presents an application of fuzzy-type methods for clustering DNA microarray data that can be applied to typical comparisons. Clustering and analyses were performed on microarray and simulated data. The results show that fuzzy-possibility c-means clustering substantially improves the findings obtained by others.

Keywords**—**Clustering, microarray data, Fuzzy-type clustering, Validation

I. Introduction

MICROARRAY technology can create an enormous amount of data quickly. It is a welcome new tool for studying such broad problems as the classification of tumors in biology and medicine. Although microarray experiments are information rich, analyzing the large amount of data obtained using this technology requires extensive data mining to identify groups of objects that share similar expression profiles [1]. For large amounts of data, clustering methods play a major role in finding groups of objects with similar functions that evince similar expression patterns of co-regulation [2], [3].

A number of methods have been proposed for clustering microarray data. Hierarchical clustering [3], [4], self-organizing maps [2], K-means [5], and fuzzy c-means [6] have all been successful in particular applications [7] and are very popular for clustering DNA microarray data. Hierarchical clustering produces dendrograms, in which each branch forms a group of objects (or genes) that have a higher-order relationship between clusters and profiles. One major shortcoming of this method is that it cannot find co-expressed objects when analyzing large amounts of microarray data, which have been collected under various biological conditions [1]. Moreover, the hierarchical clustering dendrogram is not unique and does not reflect the multiple ways in which the expression patterns of objects can be similar [8]. K-means clustering partitions data into *c* clusters such that objects in the same cluster are more similar to each other, *i.e*., the clusters are internally similar, but externally dissimilar. K-means clustering requires a fixed number of clusters based on previous knowledge of the system. These two methods result in different conclusions for the same microarray data owing to the different clustering techniques [9], [10]. The best-known clustering is c-means clustering, which minimizes the mean of square error. As the membership function of c-means is 1 or 0, it may not reflect the practical relation between the object and prototype [11]. There are two ways to tackle this drawback: fuzzy set and probability theory.

Fuzzy clustering provides a systematic, unbiased way to change precise values into several descriptors of cluster memberships [12]. These methods provide more information on the degree of similarity of each object. The main advantage of fuzzy clustering in microarray data analysis is that it explains the noise in the data. The fuzzy c-means (FCM) method [12] is one the most widely used fuzzy clustering methods for microarrays. Owing to constraints, membership cannot be translated as the typicality of the object for each cluster. FCM clustering attempts to find the most characteristic object in each cluster, which can be considered the center of the cluster, and then, the degree of membership for each object in the cluster [13]. However, the problem with FCM is that noise points can be given very similar membership in each cluster. To overcome this problem, [14] proposed possibilistic c-means clustering (PCM), which dilutes the constraint, and measures the absolute typicality of an object in a cluster. PCM clustering has the advantage of finding noise points, because a distant noisy point will belong to clusters with small possibilistic memberships. Therefore, it does not have a crucial effect on the resulting clusters. In return, PCM is very sensitive to good initialization, as well as the choice of additional parameters, and sometimes makes coincident clusters [15]. However, the memberships and typicalities are both important for correct interpretation of the data sub-structure. Pal *et al*. [15] proposed fuzzy-possibilistic c-means (FPCM) clustering, which produces both memberships and typicalities, along with the centers for each cluster.

This paper examines the application of fuzzy-type clustering to microarray data, and compares the performance of these clustering methods with existing methods. We also evaluate each of these clustering methods with validation measures for real-life and simulated datasets.

Manuscript received January 19, 2005. This work was supported by a Korea Research Foundation Grant (KRF-2002-075-C00005)

S. Y. Kim is with the Research Institute for Basic Science, Chonnam National University, Gwangju, 500-757 Korea (corresponding author to provide phone: +82-62-530-0442; fax: +82-62-530-3449; e-mail: gong@chonnam.ac.kr).

T. M. Choi is with the Department of Statistics, Chonnam National University, Gwangju, 500-757 Korea (e-mail: tmchoi@ris.chonnam.ac.kr)

II. Materials and Methods

A. Data description

Leukemia dataset: We used the leukemia dataset of [16], which consists of 38 learning samples on Affymetrix high-density oligonucleotide chips containing 7,129 human genes that are available to the public [17]. The goal of this experiment was to identify genes that were differentially expressed in 8 T-lineage acute lymphoblastic leukemia (ALL) patients, 19 B-lineage ALL patients, and 11 acute myeloid leukemia (AML) patients.

Melanoma dataset: The melanoma dataset is described in [18] and is available to the public [19]. This dataset was acquired from a study of gene expression in two types of 31 cutaneous melanomas and 7 controls. Gene expression levels were measured using cDNA microarrays containing 8,150 human genes, of which 3,613 were also identified. These ratios were transformed to a base 2 logarithmic scale.

Simulated datasets: The first dataset (simulation 1) [20] consists of two elongated clusters in two dimensions. Cluster 1 was generated by setting $x_1 = x_2 = t$, with t taking on 25 equally spaced values from -1 to 1, and adding Gaussian noise with a standard deviation of 0.1 to each variable. Cluster 2 was generated in the same way, except that the value 5 was added to each variable. The second dataset (simulation 2) consisted of three overlapping clusters in two dimensions. Two variables in each of the three clusters had bivariate normal distributions with mean vectors (0,0), (2,-2), and (-2,2), respectively, with covariance matrix Σ, where the diagonal elements are 1, and the off-diagonal elements are 0.5. In the three clusters, 25, 25, and 50 objects were generated.

B. Data pre-processing

A large number of genes exhibit nearly constant expression levels across the object and are not useful for clustering. To find significant genes, the SAM [21] method was used across multi-classes, and 5% significant genes were selected from each dataset. The 5% selection of genes resembles a real biological situation, as in [22]. The expression levels were normalized by subtracting the median and by dividing its quantile range across variable genes.

C. Fuzzy- type clustering methods

The advantage of FCM is that it always converges, while FCM searches only for the clustering solution closest to the starting center and one expects a low degree of membership for noisy points. PCM clustering was proposed to relax the constraint condition of FCM clustering [14]. PCM clustering sometimes helps when dealing with noisy data and this can sometimes be advantageous when we start with a large value of clusters and get less distinct clusters. Conversely, coincident clusters may result because the columns and rows of the typicality matrix are independent of each other. The constraint of the FCM, for an object i, makes it difficult to interpret membership as the typicality of a data point in the cluster. When estimating the cluster centers, typicality is an important means for lessening the undesirable effects of outliers [15]. Therefore, membership and typicality are both important for correct interpretation of the data substructure. FPCM clustering [15] involves defining an objective function that depends on the memberships and typicalities. The purpose of the clustering is to evolve a partition matrix **W(X)** of a given dataset, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, to find c clusters. Here, $\mathbf{x}_i$ represents the normalized expression level or $\log_2$ of the expression level of an object i. FPCM partitions **X** into c fuzzy subsets by minimizing the following objective function:

$$J_m(W,T,V) = \sum_{k=1}^{c}\sum_{i=1}^{n}(w_{ik}{}^{m} + t_{ik}{}^{\eta}) \| \mathbf{x}_i - \mathbf{v}_k \|^2 ,$$

where J_m represents the objective function defining the quality of the result obtained for prototypes **V** and membership **W**, and m is the degree of fuzzification in the clustering. A typical value of m is 2. The membership degrees w_{ik} and typicalities t_{ik} are defined such that $m,n > 1$, $0 \le w_{ik}, t_{ik} \le 1$, $\sum_{k=1}^{c} w_{ik} = 1$ for i, i=1,…,n, and $\sum_{i=1}^{n} t_{ik} = 1$ for k, k=1,…,c. η_k is used to scale t_{ik} such that $\|\mathbf{x}_i - \mathbf{v}_k\|^2 = \eta_k$. Scaling is not a necessary step in FPCM, but it may help when interpreting the typicality values [14]. $\mathbf{V} = (\mathbf{v}_k)$ is the cluster center or prototype, and $\|\mathbf{x}_i - \mathbf{v}_k\|^2$ is the Euclidean distance between each object and a fuzzy prototype.

D. Validation methods

Silhouette index: Kaufman and Rousseeuw [23] suggested selecting the number of clusters such that $k \ge 2$, 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 ($1 \le j, l \le k, j \ne l$)

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 [17]:

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

Please refer to [24] for a detailed description of the adjusted Rand index. For good clustering, we expect these values to be high. This is a useful measure when comparing two methods producing a different number of clusters.

III. Experimental Results

We conducted two experiments to examine the performance reliability of the fuzzy-type clustering methods. First, we tested cluster validation to obtain the optimal number of clusters using the silhouette index, and then, we measured the extent of agreement between two different cluster sub-structures obtained for the same set of data points using the adjusted Rand index.

A. Terminal memberships and typicalities from FPCM

Table I shows the terminal membership (w) and typicality (t) values obtained on applying FPCM to the leukemia and melanoma datasets. FPCM clustering provides a more informative description of the data than FCM alone, since it provides membership information. From the leukemia dataset, we can see that object numbers 10, 12, 30, and 31 are more equal to the membership values for each cluster, and their typicalities to the three clusters are very similar. In addition, from the melanoma data, object numbers 2, 4, 10, 24, 26, and 30 are closer to equal than the other objects in their membership values for either cluster, and their typicalities are similar. This suggests that these points are outliers. The most typical points for each of the three clusters in the leukemia data are object numbers (4,6), (17,20), and (22,23), respectively, and are object numbers 27 and 9 in either cluster of the melanoma data.

TABLE I
Terminal memberships and typicalities from FPCM

	Leukemia (c=3)						Melanoma (c=2)			
Obj	w1	w2	w3	t1	t2	t3	w1	w2	t1	t2
1	0.52	0.25	0.25	0.05	0.02	0.01	0.28	0.71	0.01	0.02
2	0.67	0.15	0.17	0.06	0.01	0.00	**0.48**	**0.51**	**0.03**	**0.02**
3	0.65	0.16	0.17	0.03	0.00	0.00	0.35	0.64	0.04	0.04
4	0.75	0.11	0.12	0.09	0.01	0.00	**0.48**	**0.51**	**0.02**	**0.01**
5	0.66	0.16	0.16	0.04	0.00	0.00	0.23	0.76	0.01	0.03
6	0.80	0.09	0.10	0.12	0.01	0.00	0.23	0.77	0.02	0.04
7	0.53	0.23	0.23	0.04	0.01	0.01	0.32	0.67	0.02	0.03
8	0.68	0.14	0.16	0.05	0.00	0.00	0.27	0.72	0.02	0.03
9	0.14	0.53	0.31	0.01	0.04	0.01	0.20	0.79	0.02	0.05
10	**0.14**	**0.43**	**0.42**	**0.01**	**0.03**	**0.02**	**0.48**	**0.51**	**0.03**	**0.02**
11	0.16	0.58	0.25	0.01	0.04	0.01	0.22	0.78	0.02	0.04
12	**0.16**	**0.42**	**0.41**	**0.01**	**0.03**	**0.02**	0.22	0.77	0.02	0.06
13	0.18	0.57	0.23	0.00	0.02	0.00	0.26	0.73	0.01	0.02
14	0.16	0.62	0.21	0.00	0.02	0.00	0.24	0.75	0.01	0.03
15	0.21	0.51	0.27	0.01	0.02	0.00	0.24	0.75	0.02	0.04
16	0.20	0.52	0.27	0.01	0.02	0.00	0.20	0.79	0.01	0.03
17	0.11	0.65	0.22	0.01	0.07	0.01	0.37	0.62	0.02	0.02
18	0.17	0.58	0.24	0.01	0.02	0.00	0.19	0.80	0.01	0.03
19	0.16	0.56	0.26	0.01	0.04	0.01	0.21	0.78	0.01	0.04
20	0.11	0.67	0.21	0.01	0.08	0.01	0.27	0.72	0.01	0.02
21	0.13	0.64	0.22	0.01	0.05	0.01	0.23	0.76	0.01	0.02
22	0.09	0.16	0.73	0.01	0.02	0.07	0.21	0.78	0.02	0.04
23	0.07	0.13	0.79	0.02	0.03	0.12	0.22	0.77	0.01	0.02
24	0.13	0.23	0.63	0.02	0.03	0.05	**0.51**	**0.49**	**0.02**	**0.01**
25	0.12	0.21	0.66	0.02	0.03	0.06	0.60	0.39	0.02	0.01
26	0.11	0.19	0.68	0.02	0.02	0.06	**0.53**	**0.47**	**0.04**	**0.02**
27	0.15	0.23	0.60	0.01	0.01	0.02	0.79	0.20	0.06	0.01
28	0.14	0.20	0.65	0.01	0.01	0.03	0.70	0.29	0.04	0.01
29	0.14	0.18	0.66	0.02	0.02	0.05	0.63	0.37	0.03	0.01
30	**0.25**	**0.31**	**0.42**	**0.03**	**0.03**	**0.02**	**0.52**	**0.47**	**0.03**	**0.02**
31	**0.19**	**0.32**	**0.48**	**0.01**	**0.01**	**0.01**	0.63	0.36	0.03	0.01
32	0.15	0.26	0.58	0.01	0.01	0.02	0.69	0.30	0.01	0.00
33	0.14	0.19	0.66	0.01	0.02	0.04	0.75	0.25	0.02	0.00
34	0.20	0.27	0.52	0.01	0.01	0.01	0.63	0.36	0.02	0.01
35	0.13	0.26	0.60	0.01	0.01	0.02	0.70	0.29	0.01	0.00
36	0.12	0.18	0.68	0.02	0.02	0.06	0.75	0.24	0.04	0.00
37	0.14	0.23	0.61	0.01	0.01	0.02	0.71	0.28	0.02	0.00
38	0.15	0.23	0.61	0.01	0.02	0.04	0.64	0.35	0.04	0.01

B. Comparative analysis

The FPCM method was also compared with hierarchical clustering using Ward's method (HC) and K-means using microarray data and simulated data. Tables I and II summarize the results of the evaluation. In general, when applied to gene expression data, FPCM and K-means clustering were able to select the correct number of clusters (Table II), and to establish the cluster membership with a high level of accuracy, as measured by the adjusted Rand index (Table III). In Simulation 1, four outliers were added to the dataset, as described in the Materials and Method. When applied to Simulation 1, all the clustering methods selected exactly two clusters as the optimal cluster, while all of the methods failed to discover the correct clusters from Simulation 2, with the overlapping clusters (Table II).

TABLE II
Estimated number of clusters using the silhouette index

	K_{true}	FPCM	HC	K-means
Leukemia	3	3	2	3
Melanoma	2	2	2	2
Simulation 1	2	2	2(or 3)	2(or 3)
Simulation 2	3	2	2	2

Although all of the methods selected the true clusters from Simulation 1 using the silhouette index, for the cases with two and three clusters, the adjusted Rand index values in Table III are equal to the values for HC and K-means, respectively. Fig. 1 shows that HC and K-means have very similar values of the Rand index with two and three clusters for the FPCM method. This means that outliers have little effect in the FPCM compared with the other methods, as shown in Fig. 1.

TABLE III
The Adjusted Rand index for FPCM, HC, and K-means. The value in parentheses is the adjusted rand index corresponding to K_{TRUE} clusters (when this differs from the estimated number of clusters)

	FPCM	HC	K-means
Leukemia	0.83	0.45 (0.75)	0.83
Melanoma	0.31	0.04	0.51
Simulation 1	0.85	0.85	0.85
Simulation 2	0.64 (0.91)	0.73 (0.81)	0.64 (0.91)

Fig. 2 shows the results using FPCM, HC, and K-means for c=2 and 3 from Simulation 1. For c=2, the results of all of the methods are equal, as shown in Fig. 2(a), but for c=3, the result using FPCM differs from the other methods, as shown in Fig. 2(b). For HC and K-means, four outlier points are classified as one cluster, while FPCM correctly does not classify outlier points as any cluster at c=3. For more detail, the membership values of the four outlier points in each clusters are (0.49,0.51), (0.49,0.51), (0.48,0.52), and (0.50,0.50), and their possibilities are closer to equal than the other points. The typicalities are also very small and similar in each cluster for other points. The two figures and Tables II and III show that FPCM is much more robust than HC and K-means for outlier points.

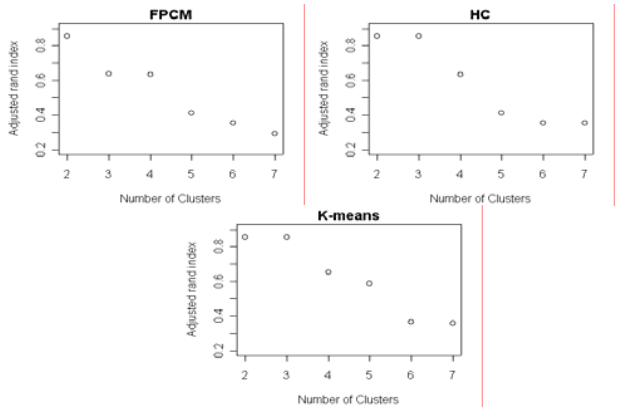
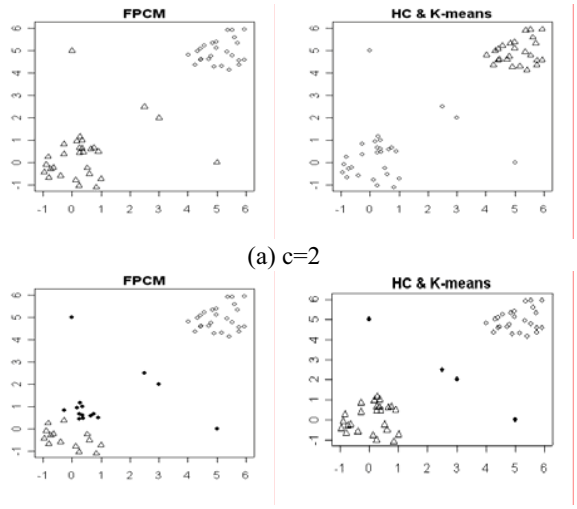


Fig. 1. Plots of the adjusted Rand index from Simulation 1.



(a) c=2

(b) c=3

Fig. 2. Clustering results for Simulation 1 at c=2 and c=3.

IV. Conclusion and Discussion

This paper examined clustering methods based on fuzzy-type, and compared the performance of fuzzy-possibilistic c-means clustering using DNA microarray data. FPCM clustering was more accurate and consistent than hierarchical clustering or the K-means method. Moreover, FPCM and K-means both represented the inherent structure of the dataset, but FPCM was superior to K-means. Unlike most hierarchical clustering methods and K-means, FPCM produces membership values, possibility values of typicality, and a set of cluster prototypes in the data matrix. Particularly, with FPCM, for a dataset with one or more large outliers, as occurs in DNA microarray data, it is possible to describe the outlier points using data examined using existing methods.

Acknowledgment

This work was also supported by Korea Research Foundation Grant (KRF-2002-075-C00005), and this work was also partly supported by grant R08-2003-000-10572-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

[1] N. Belacel, M. Cuperlovic-Culf, M.R. Boulassel, "The variable neighvorhood search metaheuristic for fuzzy clustering cDNA microarray gene expression data", Proceedings of IASTED-AIA-04 Conference. Innsbruck, Austria. February 16-18, 2004. 6 pages.

[2] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation", *Proc. Natl Acad. Sci.,* vol. 96, 1999, pp. 2907-2912.

[3] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yest Saccharomyces cerevisiae by microarray hydridization", *Mol. Biol. Cell*, vol. 9, pp. 3273-3279.

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

[5] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church, "Systematic determination of genetic network architecture", *Nat. Genet*., vol.22, 1999, pp. 281-285.

[6] R. Guthke, W. Schmidt-Heck, D. Hahn, and M. Pfaff, "Gene expression data mining for functional genomics", *Proceedings of European Symposium on Intelligent Techniques (EIST 2000)*, Aachen, Germany, 2000, pp. 170-177.

[7] S. Carla, K.H. Cho, and W. Olaf, "DNA microarray data clustering based on temporal variation: FCV with TSD preclustering", submitted 2003.

[8] B.J.T. Morgan, A.P.G. Ray, "Non-uniqueness and inversions in clusters analysis", *Applied Statistics*, vol.44, 1985, pp. 117-134.

[9] S. Chu et al., "The transcriptional program of sporulation in budding yeast", *Science*, vol.282, 1998, pp. 699-705.

[10] S. Raychaudhui, J.M. Stuart, and R.M. Altman, "Principal components analysis to summarize microarray experiments: application to sporulation time series", In *Pacific Symposium on Biocomputing*, Hawaii, 2000, pp. 452-463.

[11] J. Yu, "General c-means clustering model and it application", *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 9CVPR'03)*, 2003.

[12] J.C. Bezdek, "Pattern recognition with fuzzy objective function algorithms (New York: Plenum Press, 1981).

[13] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques", *Journal of intelligenet information system*, vol.17, 2001, pp. 107-145.

[14] R. Krishnapuram and J. Keller, "A possibilisitic approach to clustering", IEEE Trans, *Fuzzy syst*., vol. 1, 1993, pp. 98-110.

[15] N.R. Pal, K. Pal, and J.C. Bezdek, "A mixed c-means clustering model", *Fuzzy- IEEE'97*, 1997, 0-7803-3796-4.

[16] T.R. Golub, D.K. Slonim, P. Tamayo et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, vol.286, 1999, pp. 531-537.

[17] http://www.genome.wi.mit.edu/MPR.

[18] M. Bittner, P. Meltzer, Y. Chen et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling", *Nature*, vol.406, 2000, pp. 536-540.

[19] http://www.nhgri.nih.gov/DIR/Microarray/Melanoma_Supplement/index .html.

[20] S. Dudoit and J. Fridlyand, " A prediction-based resampling method for estimating the number of clusters in a dataset", Genome biology

[21] V.G. Tusher, R. Tibshirani, G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response", *Proceedings of the Natonal Academy of Science*, vol.98, 2001, pp. 5116-5121.

[22] P. Broberg, "Ranking genes with respect to differential expression", *Genome Biology*, vol.3, 2002, preprint0007.1-0007.23.

[23] L. Kaufman, P.J. Rousseeuw, "*Finding groups in data: An introduction to custer analysis*", New York: John Wiley, 1990.

[24] K.Y. Yeung, and 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.