The Korean Journal of Applied Statistics (2009)
22(1), 89–106

Descriptive and Systematic Comparison of Clustering Methods in Microarray Data Analysis

Seo Young Kim[1]

[1]Statistics Research Institute, Korea National Statistical Office

(Received August 2008; accepted October 2008)

Abstract

There have been many new advances in the development of improved clustering methods for microarray data analysis, but traditional clustering methods are still often used in genomic data analysis, which may be more due to their conceptual simplicity and their broad usability in commercial software packages than to their intrinsic merits. Thus, it is crucial to assess the performance of each existing method through a comprehensive comparative analysis so as to provide informed guidelines on choosing clustering methods. In this study, we investigated existing clustering methods applied to microarray data in various real scenarios. To this end, we focused on how the various methods differ, and why a particular method does not perform well. We applied both internal and external validation methods to the following eight clustering methods using various simulated data sets and real microarray data sets.

Keywords: Microarray, gene expression data, clustering.

1. Introduction

1.1. Clustering in microarray data

DNA microarrays are a powerful and promising biotechnology tool that enables the expression levels of thousands of genes to be monitored simultaneously. The high-throughput analysis of gene information made possible by microarrays has led to revolutionary changes in bioinformatics research (Brown and Botstein, 1999; Lander, 1999). Microarray experiments can be used to determine the genes that exhibit similar expression patterns across samples (*i.e.*, coexpressed), which may indicate a common function. Similarly, samples with similar expression profiles may share common characteristics, such as being from patients with the same disease. Identifying such groups and samples is dependent on clustering analysis of the vast amounts of microarray data generated. Many of the scenarios that can occur in microarray experiments are not supported in all existing clustering methods, which should be taken into consideration when selecting the one to use in a particular analysis. For example, a gene may be involved in more than one biological process, and hence genes included in the same subset of these processes may be represented in overlapping clusters. Moreover, genes that are not related to the samples under investigation may show a

[1]Ph.D, Statistics Research Institute, Korea National Statistical Office, Daejeon 302-120, Korea.
Email: sykim2217@nso.go.kr

nearly constant expression pattern. However, even the best clustering methods are unlikely to provide meaningful results if too much of the data are unreliable. Therefore, gene selection should form an integral part of clustering analysis.

1.2. Traditional clustering methods still predominate in microarray analyses

Clustering has many applications in studies involving microarray data. In particular, it is essential in the explanatory analysis of gene expression data (Eisen *et al.*, 1998; Golub *et al.*, 1999; Quanckenbush, 2001). Since a cluster analysis relies heavily on limited biological and medical information (*e.g.*, tumor classification), the results are not only sensitive to noise but are also prone to overfitting. There have been many advances in the development of improved clustering methods for microarray data analysis, but the use of traditional clustering methods such as agglomerative hierarchical clustering (AHC, Eisen *et al.*, 1998), DIANA (Datta and Datta, 2003), k-means (KM, Tavazoie *et al.*, 1999), partitioning around medoids (PAM, Kaufman and Rousseeuw, 1990; Dudoit and Fridlyand, 2002), self-organizing maps (SOM, Quanckenbush, 2001; Kohonen, 1997; Tamayo *et al.*, 1999), fuzzy c-means (FCM, Bezdek, 1981; Gasch and Eisen, 2002; Dembélé and Kastner, 2003), model-based clustering with a Gaussian mixture model (MBC, Yeung *et al.*, 2001; Fraley and Raftery, 2002) and bagged clustering (BAG, Leisch, 1999) predominates in genomic data analysis, which may be more due to their conceptual simplicity and their broad usability in commercial software packages than to their intrinsic merits (Handl *et al.*, 2005).

1.3. Role of a comparative study of clustering methods

Cluster analysis is a very complicated interactive process, and comparing clustering methods is not easy because formalizing the comparison as an optimization problem is highly dependent on the scenario under consideration. There were many studies on clustering methods (Yeung *et al.*, 2001; Goldstein *et al.*, 2002; Datta and Datta, 2003) that provide novel measures and study the comparative performances of their particular measures respectively. However, there are seldom studies using extensive and systematic comparisons based on various validation methods whilst considering data preprocessing strategies, class types, and data characteristics. This may be one reason why it is difficult to choose the most appropriate clustering method for a specific type of data from a microarray experiment. Instead of using the new methods that have been developed, it is commonplace for researchers to perform traditional clustering method to identify patterns in microarray data, which may not represent the best approach in a given scenario. Therefore, it is essential to assess the performances of the existing clustering methods through a comprehensive comparative study.

1.4. Goal of this study

We considered clustering of samples based on gene expression profiles. The main goal of this study was to systematically compare the predominant traditional clustering methods, and to provide some informed guidance on the behavior of each type of cluster evaluation. We focused on (1) confirming how well each method performs on a given data set, (2) finding why a particular method does not perform well and (3) determining how the different methods compare with each other. We evaluated clustering results using internal and external cluster validation measures and with several data sets, so as to determine which were more suited to particular types of data. Thus, our comparative

study did not aim to show which method is the best overall, but to show the merits and demerits of each method on the basis of various types of data, together with providing an understanding of several validations. In fact biologist suffer from selecting an optimal clustering method, and also they prefer to choose one among commom methods. Because these common methods are familiar to them, besides the methods are easy to analyze and understand on statistical software. Thus this study can be present a good gidence for selecting a clustering method for microarray data analysis.

1.5. Implementation of this study

We tested eight existing methods (AHC, FCM, PAM, KM, SOM, DIANA, BAG and MBC) using six real gene expression data sets and four types of simulated data (with three data sets of each type). We also considered data preprocessing by locally weighted scatterplot smoothing (LOWESS) normalization (Yeung *et al.*, 2001; Yang *et al.*, 2001), gene selection and missing-value imputation. For the evaluation of all clustering methods, we have used internal indices such as the silhouette index (sil, Kaufman and Rousseeuw, 1990), figure of merit (FOM, Yeung *et al.*, 2001), and the average proportion of the nonoverlap measure (NOM, Datta and Datta, 2003), and external indices such as the adjusted Rand index (aRand, Yeung and Ruzzo, 2001). As described previously, internal indices can be used to discover whether a particular method performs well, while external indices are good for assessing the performance of a method on a data set (Handl *et al.*, 2005).

2. Clustering Methods

2.1. Hierarchical clustering

Hierarchical clustering methods are used when a stratified structure of classes at different heterogeneity levels is desired. The resulting clusters can be represented as nodes of a dendrogram. The several hierarchical clustering methods differ in how they derive class distances from the dissimilarities of the objects. Two major concepts are used in implementations of hierarchical clustering techniques: divisive and agglomerative methods. It is known that hierarchical methods are more versatile in the sense that they can be more easily adapted to distance measures other than the usual distance metric. Moreover, the agglomerative method works well on data sets containing nonisotropic clusters including well-separated, chain-like, and concentric clusters, whereas a typical partitioning methods works well only on data sets having isotropic clusters (Nagy, 1968). The greatest weakness of hierarchical methods is the computational effort involved, which makes them infeasible for large data sets. In our analysis, we used both agglomerative hierarchical clustering using Wards method and divisive hierarchical clustering using DIANA (Hosel and Walcher, 2001).

2.1.1. Agglomerative hierarchical clustering Agglomerative methods generally group individual objects together to form larger and larger classes, according to the following procedure: (1) begin with n single clusters and compute the proximity matrix containing the distances between all pairs of objects, (2) find the most similar pairs of clusters using the proximity matrix and merge these pairs of clusters into single clusters and (3) update the proximity matrix to reflect this merge operation and iterate until one cluster consisting of all objects remains. Wards method assigns a homogeneity measure to every partitioning of the object set into k classes. According to the general strategy, the merging of two classes results in a new partitioning with reduced homogeneity.

2.1.2. DIANA DIANA iteratively splits the entire data set until every class consists of a single object. The object in the cluster that is most dissimilar to the other elements is then separated to

form a so-called splinter group, and the remaining elements should then be added to this splinter group.

2.2. Partitioning clustering

The partitioning methods are designed to find convex clusters in the data, such that each segment can be represented by a cluster center. The common methods PAM (Kaufman and Rousseeuw, 1990) and KM (MacQueen, 1967) are distinguished by the cluster centers in KM being averages of objects, whereas in PAM they are the actual centers. Partitioning methods work better with large data sets, since the solutions for different numbers of clusters need not be nested as in hierarchical methods. However, they are not as flexible as hierarchical methods with respect to distance measures.

2.2.1. Partitioning around medoids PAM can be regarded as a generalization of KM clustering to arbitrary dissimilarity matrices. PAM is based on the search for k representative medoids among the objects to be clustered. After finding these k medoids, k clusters are built by allocating each object to the nearest medoid, with the goal of minimizing the sum of the dissimilarities of the objects to their closest medoid. The k initial sets of medoids are first sequentially selected, and then points are swapped so as to minimize the objective function by replacing one medoid with another entry, with this step iterated until convergence. PAM is known to be more robust and computationally efficient than KM.

2.2.2. K-means KM partitions data into k clusters that 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. The algorithm first assigns each object to a cluster that has the closest centroid, and then sets the initial positions for the cluster centroids; that is, when all objects have been assigned, the positions of the k centroids are recalculated. This procedure is continued until the objects are optimally assigned to clusters (Guralnik and Karypis, 2001). In this process, KM should have little difficulty with missing data because means can still be updated and distance can still be computed.

2.3. Fuzzy c-means clustering

In non fuzzy clustering methods such as PAM, KM and AHC, each object belongs to exactly one cluster. Fuzzy clustering extends this notion to associate each object with every cluster using a membership function. For FCM, an initial fuzzy partition of the n objects into k clusters is first selected as the $n \times k$ membership matrix, where the elements of this matrix represent the grades of membership of objects in the cluster. The membership matrix is then used to find the value of a fuzzy criterion (*e.g.*, a weighted squared error criterion function) associated with the corresponding partition. FCM 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. Even though it is better than KM at avoiding local minima, FCM can still converge to local minima of the squared error criterion.

2.4. Neural network clustering

SOM is an effective method for visualizing high-dimensional data and performing clustering. SOM clustering is based on the concepts of neural networks. It converts complex, nonlinear statistical relationships between high-dimensional data into simple geometric relationships on a low-dimensional

display. The SOM network has input and output nodes. The input layer has a node for each attribute of the record, where each one is connected to every output node. Each connection is associated with a weight that determines the position of the corresponding output node. Thus as the algorithm changes the weights appropriately, the output nodes move so as to form clusters. More details on the SOM, including on the numerical procedure, are available elsewhere (Hastie *et al.*, 2001).

2.5. Bagged clustering

The basic idea of BAG is to stabilize partitioning methods such as KM by repeatedly running the cluster algorithm and combining the results (Leisch, 1999). That is, a partitioning cluster algorithm such as KM is run repeatedly on bootstrap samples from the original data. The resulting cluster centers are then combined using a hierarchical cluster algorithm. However, this method still suffers from the problem when several cluster results are obtained by repeating the procedure: there is no obvious way of choosing the best one, or combining them (Leisch, 1999).

2.6. Model-based clustering

The idea behind MBC is to regard the data as coming from a mixed distribution. MBC used in our analysis employs expectation-maximization(EM) initialized by hierarchical clustering for parameterized Gaussian mixture models (Fraley and Raftery, 2002). The two approaches are complementary: model-based hierarchical agglomeration tends to produce reasonably good partitions even when initialized without any information about the groupings, whereas initialization is critical in EM (McLachlan and Basford, 1988; Jain *et al.*, 1999) because the likelihood surface tends to have multiple modes. The number of clusters is chosen to maximize the Bayesian information criterion by initializing EM with partitions from MBC agglomeration. A detailed description of the theoretical concepts is available elsewhere (Banfield and Raftery, 1993; Fraley and Raftery, 2002).

3. Evaluation of Clusters

3.1. Adjusted figure of merit

FOM assesses the quality of clustering results by a jackknife approach (Efron and Gong, 1983). In FOM, (1) the n objects are assumed to fall into c true classes, with the ith class containing $\alpha_i n$ objects, where $0 < \alpha_i < 1$ and $\sum_{i=1}^{c} \alpha_i = 1$, (2) the expression levels of objects in class i under condition e are independent normally distributed random variables with mean $\mu_{i,e}$ and variance $\sigma_{i,e}^2$ and (3) each cluster is assumed to contain objects from only one class, and there are clusters containing class-i objects. This assumption is valid when the clustering method favors equal-sized clusters. The adjusted FOM (Yeung *et al.*, 2001) is given by

$$\mathrm{FOM}(k) = \sum_{e=1}^{m} \sqrt{\frac{1}{n} \times \sum_{i=1}^{c} (\alpha_i n - \alpha_i k)\sigma_{i,e}^2} \Big/ \sqrt{\frac{n-k}{n}}.$$

3.2. The average proportion of the nonoverlap measure

NOM computes the average proportion of objects that are not placed in the same cluster by the clustering method under consideration on the basis of the entire data set and the data sets obtained

by deleting the expression levels at a time (Datta and Datta, 2003):

$$\mathrm{NOM}(k) = \frac{1}{nl}\sum_{a=1}^{n}\sum_{i=1}^{l}\left(1 - \frac{n(C^{a,i} \cap C^{a,0})}{n(C^{a,0})}\right),$$

where $C^{a,i}$ denotes the cluster containing object a ($a = 1, \ldots, n$) in the clustering based on the data set from which the observations at variable ν_i ($i = 1, \ldots, l$) have been deleted, and $C^{a,0}$ denotes the original cluster containing object a in the clustering based on the entire data set. A good algorithm is expected to yield a small value of NOM. However, NOM is able to be affected by the numbers of objects and variables.

3.3. Adjusted rand index

The aRand computes the extent of agreement between two partitions: $C = \{C_1, C_2, \ldots, C_{k_1}\}$ is a clustering structure of a data set X, and $P = \{P_1, P_2, \ldots, P_{k_2}\}$ is a defined partition of the data. External indices of the partition agreement can be expressed in terms of a contingency table with entries n_{ij}, which are the numbers of objects of classes i and j that are members of both clusters C_j and P_j ($i = 1, 2, \ldots, k_1$, $j = 1, 2, \ldots, k_2$). Let $n_{i\cdot}$ denote the number of objects in cluster C_i (*i.e.*, row sums), and let $n_{\cdot j}$ denote the number of objects in cluster P_j (*i.e.*, column sums). The aRand can then be used to assess and compare the performance of different clustering methods (Yeung and Ruzzo, 2001). A higher value of which means a higher correspondence between two partitions:

$$\mathrm{aRand} = \frac{\sum_{ij} n_{ij}C_2 - \left(\sum_i n_{i\cdot}C_2 \sum_j n_{\cdot j}C_2\right)\Big/ {}_nC_2}{\frac{1}{2}\left(\sum_i n_{i\cdot}C_2 + \sum_j n_{\cdot j}C_2\right) - \left(\sum_i n_{i\cdot}C_2 \sum_j n_{\cdot j}C_2\right)\Big/ {}_nC_2}.$$

4. Data Sets, Preprocessing and Software

4.1. Data sets

We used both simulated and real gene expression data sets in comparisons of the eight predominant clustering methods. We employed six simulated data sets with various cluster shapes, and degrees of overlap(Table 4.1). Table 4.2 provides information on the six real gene expression data sets, briefly describing the microarray experiment applied to the various types of data set and the main features of the data sets used.

4.1.1. Simulated data sets

Sdata1. 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 $2\mathbf{I}$ covariance matrix, where the matrix $\mathbf{I}$ is an identity matrix.

Sdata2. 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.

Table 4.1. Description of five simulated datasets

Dataset	Actual clusters	Num. of dimensions	Num. of objects in each cluster	Degree of overlap among clusters
Sdata1	3	2	25, 25, 50	None
Sdata2	4	10	50, 50, 50, 50	Strong
Sdata3	2	3	100, 100	None
Sdata4	3	13	50, 50, 50	Strong
Sdata5	2	10	50, 50	Weak
Sdata6	3	70	15, 15, 15	Weak(with correlations)

Sdata3. 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.

Sdata4. 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 $\sum$, where $\sigma_{ij} = 1$, $1 \leq i \leq 3$ and $\sigma_{ij} = 0.5$, $1 \leq i \neq j \leq 3$. The remaining 10 noise variables are generated independently from the $N_{10}(\mathbf{0}, \mathbf{I})$ distribution. Each cluster contains 50 objects.

Sdata5. 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.

Sdata6. Three clusters with interactions between variables. This clusters induced by variables 1 to 20 and 21 to 40 variables that are generated to have different expression profiles across objects; for the first 20 variables, $i = 1, \ldots, 20$, each cluster is chosen to have 15 objects from $N(0, 1)$, $N(3, 1)$ and $N(-3, 1)$, and the other 20 variables, $i = 21, \ldots, 40$, each cluster is chosen to have 15 objects from $N(0, 1)$ and $N(-3, 1)$ and $N(3, 1)$. The remaining 30 noise variables are generated as $N_{30}(\mathbf{0}, \mathbf{I})$.

Note that above configurations in generating the datasets were considered in earlier papers (Kaufman and Rousseeuw, 1990; Fraley and Raftery, 2002; Ilana, 2006). The configurations are summarized in Table 4.1.

4.1.2. Microarray data The clustering methods described in section of Clustering method were applied to gene expression data from six published cancer microarray studies. The following Colon and Lymp were used as the representative heterogeneous data sets, and Lung and NCI were used as the large samples.

Leukemia. The Leuk data set consisted of 72 samples on the Affymetrix high-density oligonucleotide chips containing 3571 human genes (Golub *et al.*, 1999). The data comprised 47 ALL patients (38 ALL B-cells, 9 ALL T-cells) and 25 AML patients. These data were obtained after performing preprocessing as described previously (Dudoit and Fridlyand, 2002; Lee *et al.*, 2005).

Melanoma. The Mela data set consisted of 38 samples from both tissue biopsies and tumor cell lines: 31 cutaneous melanomas and 7 controls (Bittner *et al.*, 2000). Gene expression levels

Table 4.2. Description of real gene expression data sets

Data	Num. of class	Components of classes	Num. of samples	Num. of genes	Chip type
Leuk	$k = 3$	AML (25), ALL B-cells (38), ALL T-cells (9)	72	3571	Affymetrix
Mela	$k = 2$	Group 1 (31), Group 2 (7)	38	3613	cDNA
Colon	$k = 2$	Tumor tissue (40), normal colon (22)	62	2000	Affymetrix
Lymp	$k = 3$	B-cells chronic (11), FL (9), DLBCL (42)	62	4026	cDNA
Lung	$k = 5$	AD (139), SQ (21), COID (20), SCLC (6), NL (17)	203	12600	Affymetrix
NCI	$k = 8$	Breast cancer (7), CNS (6), colon cancer (7), leukemia (6), melanoma (8), NSCLC (9), ovarian cancer (6), renal cancer (8), prostate cancer (2)	57	8150	cDNA

were measured using cDNA microarrays containing 8150 human genes. The filtering method excluded genes with expression ratios greater than 50 and less than 0.02 (Darlene *et al.*, 2002), which resulted in 3613 of the 8150 genes being used in the analysis.

Lung cancer. The Lung data set has been described previously (Bhattacharjee *et al.*, 2001). This data set came from a study of gene expression in five types of lung carcinoma: 139 lung AD, 21 SQ, 20 COID, 6 small-cell lung carcinomas cases and 17 NL. Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing 12000 human genes.

Colon cancer. The Colon data set consisted of both 40 normal samples and 22 tumor samples on the Affymetrix oligonucleotide chips containing 6500 human genes (Alon *et al.*, 1999), of which we used 2000 genes of the 62 samples as described previously (Alon *et al.*, 1999).

Lymphoma. The Lymp data set came from a study of cDNA gene expression, and these data were composed of 92 samples for the three most prevalent adult lymphoid malignancies: B-CLL, FL and DLBCL (Alizadeh *et al.*, 2000). We used 62 samples and 4026 genes, including 11 of B-CLL, 9 of FL and 41 of DLBCL as described previously (Lee *et al.*, 2005).

NCI60. The NCI data set consisted of 60 cell lines (Ross *et al.*, 2000) derived from tumors with different sites of origin: 7 breast cancer, 6 central nervous system, 7 colon cancer, 6 leukemia, 8 melanoma, 9 non-small-cell lung carcinoma, 6 ovarian cancer, 8 renal cancer, 2 prostate cancer and 1 unknown. We first selected 8150 genes that had at most three missing values and used 57 samples excluding two classes of 2 prostate cancer and 1 unknown classes with small samples.

4.2. Data preprocessing and software

4.2.1. Missing value imputation For the Lymp and NCI data sets, missing values were imputed by the k-nearest-neighbor algorithm using $k = 5$ neighbors to estimate the missing value, with the selection of the neighbors based on sample correction (Troyanskaya *et al.*, 2001).

4.2.2. Normalization The samples were normalized by LOWESS across genes as described previously (Yeung *et al.*, 2001). The main idea of LOWESS involves obtaining the calibration factor

Table 4.3. R packages and functions corresponding to clustering, validation, missing value imputation, and gene selection method used in the study.

	Methods	Package	Function
Clustering	AHC(Agglomerative hierarchical clustering)	stats	hclust
	FCM(Fuzzy c-means)	e1071	cmeans
	PAM(Partitioning around medoid)	cluster	pam
	KM(K-means)	stats	kmeans
	SOM(Self-organizing map)	som	som
	DIANA(Diana)	cluster	diana
	BAG(Bagged clustering)	e1071	bclust
	MBC(Model-based clustering)	mclust	Mclust
Validation	sil(Silhouette)	cluster	silhouette
	FOM(Figure of merit)	SAGx	fom
	NOM(Non-overlap measure)	our code	using R
	aRand(Adjusted Rand index)	e1071	classAgreement
Missing value	K-nearest neighbor	class	knn
Gene selection	Coefficient of variation	our code	using R

using a locally weighted polynomial regression of the intensity scatterplot. It is practically reasonable to apply different preprocessing method by cDNA and Affymetrix microarrays. Intensity values measured by Affymetrix microarray can be normalized by various preprocessing methods, such as dChip, GCRMA, RMA and MAS. The previous study presented that, on practical applications of microarray-based research, the choice of preprocessing method is of minor influence on the final analysis outcome of large microarray studies (Verhaak *et al.*, 2006). Also this study is focusing to compare the clustering method and thus author is going to apply just LOWESS normalization method on convience of data handling procedure. Compared to other methods, the LOWESS method is known to be robust across a wider range of types of data sets. Note that microarray data need to be normalized so as to remove systematic variations in the experiments that affect the measured expression levels.

4.2.3. Variable selection Expression levels were measured for thousands of genes in each of the data sets. Many of the genes exhibited nearly constant expression levels, as indicated by the coefficient of variation across the objects (Dudoit and Fridlyand, 2002). These genes did not seem to be useful for classification purposes; therefore, we only used genes with a high coefficient of variation of expression levels across classes from the clustering process. We selected the 100 genes with the highest coefficients of variation from the Leuk, Mela, Lymp, and Colon data sets, and 200 genes from Lung and NCI data sets (since they contain more classes). Because an arbitrary selection method was used, as described previously (Dudoit and Fridlyand, 2002) we also tested the effects of increasing the number of genes to 300~500 or decreasing the number to less than 100; this had no significant effect on the results.

4.2.4. Software and availability All simulation and analyses were carried out with version 2.3 of R (http://www.R-project.org) using packages stats, e1071, cluster, som, mclust, SAGx and class. For NOM and gene selection, we have directly written the source code of them by R language. Table 4.3 summarizes the package and function corresponding to each method used in our study. And R codes and data sets used in the study are available from the authors on requests.

Table 5.1. Estimating the number of clusters by applying the eight methods to the simulated data sets

Data set	Actual	FCM	PAM	KM	SOM	AHC	DIANA	BAG	MBC
Sdata1	3	3	3	3	3	3	3	3	3
Sdata2	4	4	3	2	4	2	3	2	2
Sdata3	2	2	2	2	2	2	2	2	2
Sdata4	3	2	2	2	2	2	5	2	2
Sdata5	2	2	2	2	2	2	2	2	2
Sdata6	3	3	3	2	3	2	3	2	2

5. Results and Discussion

The focus of our study was the descriptive and comprehensive comparison of microarray data clustering. We applied eight existing cluster methods to both real and simulated data sets. For each data set, evaluations were conducted using three internal cluster indices (sil, FOM and NOM) and an external cluster index (aRand). Most of the internal indices can be used to estimate the number of clusters in a data set, which commonly includes the computation of clustering results for different numbers of clusters, with the internal indices also varying with the number of clusters (Handl *et al.*, 2005). We adopted the sil to determine the number of clusters in our study in terms that this method was often used in other studies (Dudoit and Fridlyand, 2002; Handl *et al.*, 2005) as a good measure. The sil is known to be preferable over other noise-sensitive methods, given the noisy nature of data (Handl *et al.*, 2005). We also checked the characteristics of clustering methods regularly using both internal and external indices, because their results differ with the clustering method. We first used simulated data sets with four types of data (as described in the Methods section) where the correct solution was known a priori, to assess the practical ability of traditional methods to find natural clusters and examine the clustering properties of the used methods. We then assessed the performance of each clustering method in terms of internal and external measures, and inspected the appropriateness of each method when it was applied to real gene expression.

5.1. Simulated data analysis

For the simulated data analysis, Table 5.1 displays the estimated number of clusters for the sil, in which the number of clusters was selected based on the largest value of sil for values computed from 2 to 10. Table 5.2 compares the results for the FOM, NOM and aRand.

5.1.1. Clustering evaluation in terms of internal measures Overall, all methods successfully identified the true number of clusters in Sdata1, Sdata3 and Sdata5 with weaker overlapping clusters. For all data types except for Sdata4, SOM and FCM correctly determined the true number of clusters. These results indicate that SOM and FCM exhibit good overall performance in estimating the true number of clusters from simulated data. PAM and DIANA also exhibits good performance, although this is somewhat dependent on the degree of overlapping due to heterogeneity between clusters. In contrast, KM, AHC, BAG and MBC might not be appropriate for identifying the true number of clusters.

And, one observation is that stronger overlap between clusters result in a lower prediction power for FOM. Table 5.2 indicates that the values of FOM tended to increase with increasing the degree of overlap between clusters. In particular, PAM has a good performance for Sdata2 and Sdata1 with strong overlapping clusters. On the other hands, for Sdata6 (with corrleation between variables), BAG and MBC produced good clustering results in terms of the internal FOM index. BAG and

Table 5.2. Comparison results with the true clusters when applying the eight clustering methods to simulated data sets.

	FCM	PAM	KM	SOM	AHC	DIANA	BAG	MBC
			FOM					
Sdata1	46.58	42.68	42.64	42.20	42.64	28.10	43.22	43.41
Sdata2	72.31	67.37	89.90	90.49	89.90	83.99	83.99	54.62
Sdata3	11.94	11.94	11.94	11.94	11.94	11.94	11.94	11.94
Sdata4	73.69	49.98	60.49	47.97	60.49	61.61	61.61	71.39
Sdata5	45.49	48.13	37.15	45.91	37.15	48.04	48.04	56.12
Sdata6	23.61	23.29	22.95	20.37	22.73	22.44	15.19	17.99
			NOM					
Sdata1	0	0	0.16	0	0	0	0.05	0
Sdata2	0	0	0.17	0	0	0	0.21	0
Sdata3	0	0	0	0	0	0	0	0
Sdata4	0	0	0.08	0	0	0	0.34	0
Sdata5	0	0	0	0	0	0	0.23	0
Sdata6	0	0	0.13	0	0	0	0.27	0
			aRand					
Sdata1	1	1	1	0.73	1	0.94	0.98	1
Sdata2	0.58	0.54	0.50	0.35	0.45	0.46	0.47	0.74
Sdata3	1	1	1	1	1	1	1	1
Sdata4	0.30	0.30	0.33	0.26	0.21	0.22	0.29	0.28
Sdata5	0.57	0.49	0.57	0.51	0.57	0.60	0.22	0.67
Sdata6	0.92	0.88	0.41	0.79	0.86	0.90	0.79	0.93

KM showed a low consistency from NOM of leave-one-out validation. Therefore, the results from BAG and KM might depend on the specific genes that are used in the analysis.

5.1.2. Clustering evaluation in terms of external measures The aRand was used to assess the accuracy of the implemented clustering methods. The aRand preserves information on the consistency of different clustering of the same data, and has been considered a good metric for clustering evaluations (Milligan and Cooper, 1986; Monti *et al.*, 2003).

As shown in Table 5.2, in the external validation, all the clustering methods except for SOM tended to exhibit similar performances for Sdata1 and Sdata3, which had clusters that overlapped less due to the small variances. In contrast, for Sdata2 and Sdata4 with strong overlapping clusters, FCM, PAM and MBC showed a high consistency rate, especially for Sdata6 with heterogeneity between variables, and FCM, DIANA and MBC outperformed the other methods. FCM and PAM generally give stable results that were also the best for all data sets. Moreover, it is interesting that PAM and DIANA - which are conceptually different approaches - exhibit similar performance, while the performances of SOM and KM differed from that of PAM, despite them using conceptually the same approach.

5.2. Microarray data analysis

We evaluated the performance of applying the eight predominant traditional clustering methods to six real data sets. For a comprehensive comparison, we focused on interpreting the clustering results using two parameters: the class size and the heterogeneity of the data. We searched for characteristics of the clustering methods using descriptive means on the basis of real cluster struc-

Table 5.3. Number of clusters from microarray data sets by the sil

	Actual	FCM	PAM	KM	SOM	AHC	DIANA	BAG	MBC
Leuk	3	3	2	3	2	3	2	2	2
Mela	2	2	2	2	2	2	2	2	2
Lymp	3	2	2	2	2	2	2	2	2
Colon	2	4	4	4	4	4	4	2	2
Lung	5	3	2	3	2	2	2	2	2
NCI	8	2	2	2	2	2	2	2	2

Table 5.4. Comparison results corresponding to the known clusters when applying the clustering methods to the six microarray data sets.

	FCM	PAM	KM	SOM	AHC	DIANA	BAG	MBC
			FOM					
Leuk	18.09	19.95	19.09	29.43	19.23	22.66	34.73	68.07
Mela	9.05	8.86	9.05	8.98	10.02	9.05	11.83	23.28
Lymp	17.10	21.76	23.72	18.36	24.33	29.38	31.02	70.61
Colon	33.91	33.91	33.95	33.92	33.75	33.92	58.43	109.35
Lung	3048.1	2960.1	3042.2	2495.1	2828.3	4760.2	2505.1	9300.2
NCI	6.01	5.21	6.22	4.60	6.41	7.31	9.97	57.05
			NOM					
Leuk	0	0	0.13	0	0	0	0.30	0
Mela	0	0	0	0	0	0	0.16	0
Lymp	0.09	0	0.01	0	0	0	0.08	0
Colon	0	0	0	0	0	0	0.05	0
Lung	0.02	0	0.14	0	0	0	0.34	0
NCI	0.30	0	0.26	0	0	0	0.33	0
			aRand					
Leuk	0.60	0.60	0.59	0.53	0.65	0.53	0.59	0.63
Mela	0.31	0.33	0.32	0.35	0.31	0.31	0.40	0.31
Lymp	0.41	0.35	0.47	0.35	0.41	0.71	0.69	0.41
Colon	0.49	0.49	0.49	0.47	0.50	0.49	0.53	0.49
Lung	0.57	0.53	0.58	0.36	0.38	0.68	0.39	0.38
NCI	0.39	0.24	0.40	0.21	0.24	0.28	0.25	0.35

tures. For real data, we first selected genes considered useful for clustering, with a high variance in the expression levels across samples as mentioned in the Methods section (Dudoit and Fridlyand, 2002). Table 5.3 lists the estimated number of clusters by the sil, and Table 5.4 lists the values of validation indices for the FOM, NOM and aRand for real data sets.

5.2.1. Estimated number of clusters In the Leuk, Mela, Lymp and Colon data sets, the esistence of three, two, two and three classes that were well known a priori, repectively. For these data sets, all of the clustering methods evaluated by the sil identified exactly two clusters for the Mela data set, while all methods misclassified the three clusters for the Lymp data set as two clusters. Moreover, for Colon data set all clustering methods excepting BAG and MBC, which identified exactly two clusters, misclassified the actual two clusters as four clusters. For the Leuk data set, FCM, KM and AHC correctly determined the actual three clusters, whereas the other clustering methods determined that there were only two clusters. This result is not surprising because many studies have found only two clusters, corresponding to acute lymphoblastic leukemia(ALL) and

acute myeloid leukemia(AML), even though three classes corresponding to the ALL T-cell, ALL B-cell, and AML samples are clearly evident in the correlation matrix. For the Lung and NCI data sets with large clusters, all clustering method did not find the actual clusters.

5.2.2. Performance of clustering methods based on heterogeneous data Especially interesting results were obtained for the clustering of the Lymp and Colon data sets. As mentioned previously (Milligan and Cooper, 1986; Lee *et al.*, 2005), the variability in expression may differ between clusters. In the Lymp data set, the average expression of follicular lymphoma(FL) and B-cell chronic lymphocytic leukemia(B-CLL) subclasses has been shown to be much more variable than that of diffuse large-B-cell lymphoma (DLBCL; Tibshirani *et al.*, 2003).

Therefore, in our analysis both Colon and Lymp data sets could be considered as the examples of heterogeneous data. In these heterogeneous data sets, most clustering methods (except for BAG and MBC in the Colon data set) failed to find the actual clusters as shown in Table 5.3. In particular, when all clustering methods were applied to the Lymp data set, one cluster consisted of FL and DLBCL and the other consisted of CLL. This result is consistent with that obtained using only PAM method (Dudoit and Fridlyand, 2002). From this result, we supposed that CLL samples were obtained from peripheral blood cells, as opposed to lymph-node biopsy specimens for the FL and DLBCL samples. Moreover, for the Colon data set, the correlation for the most variable genes suggest the existence of a subclass of tumors (Dudoit and Fridlyand, 2002). Indeed, applying all the clustering methods to the Colon data set discriminated 44 samples from the tumor groups. Thus, we would expect to identify at most two classes for this data set.

5.2.3. Performance of clustering methods based on data with large clusters For the lung cancer(Lung) and NCI60(NCI) data sets with large classes, all the clustering methods appear to underestimate the number of clusters (Table 5.3). In fact, for these data sets the classes are not clearly distinguishable. The Lung data set is highly skewed to the sample of a class; that is, the adenocarcinoma(AD) class includes 139 of all the 203 samples (Table 4.2). Therefore, in the image of the correlation matrix, only two or three classes tend to cluster together (Table 5.3). Although the clustering results differ between the methods, the two AD and squamous cell lung carcinomas(SQ)/pulmonary carcinoids(COID) classes were identified by MBC and DIANA, and the three AD, COID, and SQ/normal lung specimens(NL) class were identified by KM and PAM. The performances of KM and PAM differed slightly even though they use a similar clustering approach to partition the samples. As shown previously (Dudoit and Fridlyand, 2002; Tibshirani *et al.*, 2003), NCI data set has a complicated structure, in that it includes eight cancer classes, with colon and leukemia classes exhibiting strong interclass correlations, and each class is very small, which makes it difficult to distinguish classes. Therefore, when using such complicated data, we should not expect to identify exactly the known number of clusters, and we also found that the results of clustering are strongly influenced by the data structures.

5.2.4. Validations of internal and external clusters Table 5.4 lists the evaluation result for the internal cluster indices(FOM and NOM) and the external cluster index(aRand). First, for the Leuk and Mela data sets, which have a more distinct cluster structure, FCM, PAM and KM (which use a partitioning algorithm) showed good performances based on the FOM, with the best performance being for the aRand. While BAG and MBC are the worst based on the FOM, these two methods are comparable with the remaining methods based on the aRand. Second, for the Lymp and Colon data sets, which exhibit a heterogeneous structure between clusters, FCM, PAM and SOM are superior to other methods based on the FOM, but DIANA, BAG and KM outperformed the other methods based on the aRand. This reflects that results from assessments of clustering methods differ

significantly between internal and external cluster validations. Finally, for the Lung and NCI data sets with large classes, SOM, PAM and AHC could well perform based on the FOM, while DIANA, KM and FCM are superior to the other methods based on the aRand. Applying PAM to the Lung data set and MBC to the NCI data set (with the largest classes) resulted in reasonable performance based on the aRand. On the other hand, the results of BAG and KM clustering were generally inconsistent, similar to the simulated results based on the NOM. However, for the Lymp and Colon data sets, BAG and KM clustering methods exhibited relatively high consistency respectively as in evaluations using the aRand, which suggests that BAG and KM tend to discriminate clusters well from heterogeneous data set.

5.2.5. Attention and emphasis prior to selecting the clustering method The above comparisons demonstrate that the performances of the same clustering method in evaluations differ between the use of internal and external clusters indices in all data set analyses. Thus, special attention is necessary when a developer is evaluating a novel method and a user is selecting an existing method: how strong the connectedness of samples will be within clusters, how distinguishable the samples are from the clusters, what characteristics could be in the data, and how many samples could be overlapped in cluster. The above factors are considered to some extent even though clustering method basically uses an unsupervised algorithm. That is, in microarray experiments, researchers already often have ideas about the various subgroups of interesting expression patterns that are to be expected for sample clustering of a particular data set. Therefore, careful inspection of the data could play an important role in choosing the most appropriate clustering method - which is a common descriptive tool prior to core analysis - in complicated data analyses such as those involving microarray data.

6. Conclusions

This study compared the eight predominant traditional clustering methods in terms of their abilities to identify subgroups of samples that express similar patterns. The clustering results differed greatly with the clustering method that was applied. The choice of the most appropriate method is often confusing in practical applications, with clear guidelines not yet being available. One of our aims was therefore to present some guidelines for the choice of clustering method favorable for applying to microarray data. For this purpose, internal and external cluster validation methods were applied to practical data sets. We analyzed simulated data sets containing different types of cluster structures, as well as real data sets representing practical scenarios. The comparison results and our recommendations can be summarized as follows:

- Cluster analysis provides clues to the function of unknown samples based on comparisons with functions of known coregulated samples (Grotkjaer *et al.*, 2006). Evaluating clustering using a single validation criterion may not result in the most appropriate clustering method being selected. Moreover, choosing the best clustering method is problematic due to different evaluations of the same clustering method yielding different results. This study has shown that most internal indices suffer from biases with regard to the number of clusters, and they might also exhibits biases with regard to the shape of the underlying data manifold and the clustering structure. External indices suffer from biases with respect to the number of clusters, the distribution of cluster sizes, and the actual number of clusters (Halkidi *et al.*, 2001). A previous comparative study (Dudoit and Fridlyand, 2002) showed that DIANA generally produces good results, but this could be misleading given that the performance was assessed

using only internal validation. Thus, obtaining reliable information from data requires careful interpretation of clustering results, and the results should possibly also be confirmed by multiple validation methods incorporating internal and external cluster validations due to many validation methods exhibiting bias.

- The performance of clustering methods varies significantly between assessments by internal and external validation. In general, when applied to simulated data sets, SOM, FCM and PAM perform well in identifying the number of clusters using internal validation. For real data sets, FCM and PAM perform well compared to other methods with homogeneous data sets, while FCM, PAM and SOM perform better with heterogeneous data sets. Overall, PAM and FCM present similarly good results by internal validation. BAG and MBC always perform worse for real data sets based on only FOM. In contrast, FCM, PAM and MBC perform better compared to the other methods by external evaluation based on the aRand. In particular, BAG and MBC are comparable to other methods for data with small clusters, with BAG performing especially well with heterogeneous data sets in external validation. However, care is needed when using FCM and SOM, which require several parameters to be specified in advance, and this disadvantage restricts their applicability to microarray data analysis.
- BAG and KM perform better on data with heterogeneity and small classes, in terms of the external consistency, whereas they are somewhat unstable in clustering data according to the selected data point or the variables. This might be due to BAG constructing many bootstrap samples by drawing with replacement from the original data set, and then running KM as the base method on each resampling set (Leisch, 1999). That is, the centers depend on the inputted samples, and KM itself find only a local minimum of the error function after many runs. PAM, which uses a similar partitioning algorithm, can be an appropriate alternative in that it is more robust in choosing the number of clusters and the initial center of a cluster, and in assigning the sample to each cluster than KM, particularly when the true number of clusters is unknown.
- It is not surprising that there is no single choice for the best clustering method. It is our view that the various clustering methods work differently (1) in estimating the correct number of clusters and (2) in allocating the samples to clusters. Here we recommend incorporating these two features as a compromise solution: that is, SOM, PAM and FCM appear to be solid and robust performers under internal validation, whereas DIANA, FCM and PAM perform well in assigning the samples to clusters.

In fact this study focused to compare traditional clustering methods and thus other novel methods did not compare. However, according to the aim of research, many other interesting methods can be applicable to cluster microarray data. For example the tight clustering method is to find a cluster the most informative, tight and stable clusters of sizes (Tseng and Wong, 2005). In other words, most traditional clustering algorithms aim to assign all genes into clusters, but tight clustering focuses to identify the core patterns and the result becomes more interpretable. Also, bi-clustering method is conceptual clustering approach within categorical data. This method provides a collection of bi-clusters, *i.e.*, linked clusters for both objects and attribute value pairs (Pensa and Boulicant, 2005). These methods could work well in some research with special aim and experimental conditions. The extensive comparison and review for novel clustering methods may help user to choose appropriate method to the data.

On the other hands, it is often the case that biologists have prior ideas of what constitutes a good choice for the number of clusters, at least approximately. For good clustering, it is essential to not only have a thorough understanding of the particular tool being used, but also to know the details of the data gathering process and to have some domain expertise. The more information the user has about the data at hand, the greater the likelihood of success in assigning its true class structure (Jain *et al.*, 1999). Despite many difficulties in comparative testing, our results could be useful to the optimal clustering of microarray data. This study suggests that the most appropriate clustering method should be selected according to the particular biological goal. Also, this study may present a good guidence to a non-statician for clustering microarray data under consideration of various real situation. However, the use of clustering methods to search for matches remains a significant problem that deserves further investigation.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503–511.

Alon, U., BarKai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceeding of the National Academy of Sciences*, **96**, 6745–6750.

Banfield, J. D., Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering, *Biometrics*, **49**, 803–822.

Bezdek, J. C. (1981). *Pattern Rcognition with Fuzzy Objective Function Algorithms*, Plenum press, New York.

Bhattacharjee, A., Richards, W. G., Staunton, J. Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub-classes, *Proceeding of the National Academy of Sciences*, **98**, 13790–13795.

Bittner, M., Meltzer, P. and Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N. and Trent, J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature*, **406**, 536–540.

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays, *The Chipping Forecast*, **21**, 33–37.

Darlene, R. G., Debashis, G. and Erin, M. C. (2002). Statistical issues in the clustering of gene expression data, *Statistica Sinica*, **12**, 219–240.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data, *Bioinformatics*, **19**, 459–466.

Dembélé, D. and Kastner, P (2003). Fuzzy C-means method for clustering microarray data, *Bioinformatics*, **19**, 973–980.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biology*, **3**, research0036.1–0036.21.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *American Statistician*, **37**, 36–48.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceeding of the National Academy of Sciences*, **95**, 14863–14868.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.

Gasch, A. P. and Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, *Genome Biology*, **3**, research0059.

Goldstein, D. R., Ghosh, D. and Conlon, E. M. (2002). Statistical issues in the clustering of gene expression data, *Statistica Sinica*, **12**, 219–240.

Golub, T. R., Slonim, D. K. and Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531–537.

Grotkjaer, T., Winther, O., Regenberg, B., Nielsen, J. and Hansen, L. K. (2006). Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm. *Bioinformatics*, **22**, 58–67.

Guralnik, V. and Karypis, G. (2001). A scalable algorithm for clustering protein sequences, *In Workshop on Data Mining in Bioinformatics, Proceedings of the U.S.A.*, 73–80.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques, *Journal of Intelligenet Information System*, **17**, 107–145.

Handl, J., Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis, *Bioinformatics*, **21**, 3201–3212.

Hastie, T., Tibshirani, R. and Fredman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York.

Hosel, V. and Walcher, S. (2001). Clustering techniques: A brief survey, Technical Report, *Institute of Biomathematics and Biometry.*

Ilana, B.-L. (2006). A generalized clustering problem, with application to DNA microarrays, *Statistical Applications in Genetics and Molecular Biology*, **5**, Article 2.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*, Prentice-Hall, Inc., Upper Saddle River, New Jersey.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: A Review. *ACM Computing Surveys*, **31**, 264–323.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, New York.

Kohonen, T. (1997). *Self-Organizing Maps*, Springer, Heidelberg.

Lander, E. S. (1999). Array of hope, *Nature Genetics*, **21**, 3–4.

Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis*, **48**, 869–885.

Leisch, F. (1999). Bagged clustering. Working Paper Serise 51, SFB, *Adaptive Information Systems and Modeling in Economics and Management Science.*

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the 5th Berkely Symposium*, **1**, 281–297.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: inference and applications to clustering*, Marcel Dekker, New York.

Milligan, G. W. and Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavioral Research*, **21**, 441–458.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, *Machine Learning Journal*, **52**, 91–118.

Nagy, G. (1968). State of the art in pattern recognition, *Proceedings of the IEEE*, **56**, 836–862.

Pensa, R. G., Robardet, C. and Boulicaut, J.-F. (2005). *LNAI* **3721**, 643–650.

Quanckenbush, J. (2001). Computational analysis of microarray data, *Nature Review Genetics*, **2**, 418–427.

R Development Core Team. R: A language and environment for statistical computing. 2004 [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, Austria [ISBN 3-900051-00-3].

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, **24**, 227–234.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceeding of the National Academy of Science*, **96**, 2907–2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999). Systematic determination of genetic network architecture, *Nature Genetics*, **22**, 281–285.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**, 104–117.

Troyanskaya, O., Cantor, M., Sherlock, G. Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525.

Tseng, G. C. and Wong, W. H. (2005). Tight clustering: A Resamping-based approach for identifyng stable and tight patterns in data, *Biometrics*, **61**, 10–16.

Verhaak, R. G. W., Staal, F. J. T., Valk, P. J. M., Lowenberg, B., Reinders, M. J. and de Ridder, D. (2006). The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies, *BMC Bioinformatics*, **7**, 105.

Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. (2001). Normalization for cDNA microarray data, *Optical Technologies and Informatics*, **42**, 141–152.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data, *Bioinformatics*, **17**, 977–987.

Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001). Validating clustering for gene expression data, *Bioinformatics*, **17**, 309–318.

Yeung, K. Y. and Ruzzo, W. L. (2001). An empirical study on principal component analysis for clustering gene expression data, *Bioinformatics*, **17**, 763–774.