Chapter 16

CLUSTERING OR AUTOMATIC CLASS DISCOVERY: NON-HIERARCHICAL, NON-SOM

Clustering algorithms and assessment of clustering results

Ka Yee Yeung

Department of Microbiology, University of Washington, Seattle, WA 98195, USA. e-mail: kayee@cs.washington.edu

1. Introduction

DNA microarrays offer a global view on the levels of activity of many genes simultaneously. In a typical gene expression data set, the number of genes is usually much larger than the number of experiments. Even a simple organism like yeast has approximately six thousand genes. It is estimated that humans have approximately thirty thousand to forty thousand genes (Lander et al., 2001).

The goal of cluster analysis is to assign objects to clusters such that objects in the same cluster are more similar to each other while objects in different clusters are as dissimilar as possible. Clustering is a very well-studied problem, and there are many algorithms for cluster analysis in the literature. Please refer to (Anderberg, 1973), (Jain and Dubes, 1988), (Kaufman and Rousseeuw, 1990), (Hartigan, 1975) and (Everitt, 1993) for a review of the clustering literature. Because of the large number of genes and the complexity of biological networks, clustering is a useful exploratory technique for analysis of gene expression data.

In this chapter, we will examine a few clustering algorithms that have been applied to gene expression data, including Cluster Affinity Search Technique (CAST) (Ben-Dor and Yakhini, 1999), (Ben-Dor et al., 1999), k-means (MacQueen, 1965), (Tavazoie et al., 1999), Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 1990), and model-based clustering (Fraley and Raftery, 1998), (Yeung et al., 2001a).

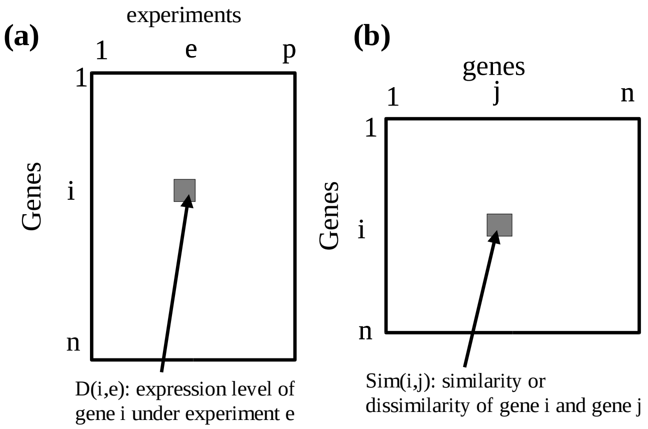
(a)
experiments
1
e
p
1
Genes
i
n
D(i,e): expression level of gene i under experiment e
(b)
genes
1
j
n
1
Genes
i
n
Sim(i,j): similarity or dissimilarity of gene i and gene j

Figure 16.1. (a) A data matrix. (b) A similarity matrix.

2. Background and Notations

A data set containing objects to be clustered is usually represented in one of two formats: the *data matrix* or the *similarity (or dissimilarity) matrix*. In a data matrix, the rows usually represent objects, and the columns usually represent features or attributes of the objects. Suppose there are n objects and p attributes. We assume the rows represent genes and the columns represent experiments, such that entry (i, e) in the data matrix D represents the expression level of gene i under experiment e, where $1 \leq i \leq n$ and $1 \leq e \leq p$ (see Figure 16.1a). The ith row in the data matrix D (where $1 \leq i \leq n$), D_i, represents the *expression vector* of gene i across all p experiments. In clustering genes, the objects to be clustered are the genes. The similarity (or dissimilarity) matrix contains the pairwise similarity (or dissimilarity) of genes. Specifically, entry (i, j) in the similarity (or dissimilarity) matrix Sim represents the similarity (or dissimilarity) of gene i and gene j, where $1 \leq i, j \leq n$ (see Figure 16.1b). The similarity (or dissimilarity) of gene i and gene j can be computed using the expression vectors of gene i and gene j from the data matrix. Hence, the similarity (or dissimilarity) matrix Sim can be computed from the data matrix D. For the rest of the chapter, the objects to be clustered are the genes in a given gene expression data set unless otherwise stated.

3. Similarity metrics

The measure used to compute similarity or dissimilarity between a pair of objects is called a *similarity metric*. Many different similarity metrics have been used in clustering gene expression data, among which the two most popular similarity metrics are *correlation coefficient* and

Euclidean distance. Correlation coefficient is a similarity measure (a high correlation coefficient implies high similarity) while Euclidean distance is a dissimilarity measure (a high Euclidean distance implies low similarity).

The correlation coefficient between a pair of genes i and j $(1 \leq i, j \leq n)$ is defined as

$$\sum_{e=1}^{p} \frac{(D(i,e) - \mu_i)(D(j,e) - \mu_j)}{(\parallel D_i \parallel \parallel D_j \parallel)} \tag{16.1}$$

where $\mu_i = \sum_{e=1}^{p} D(i,e)/p$ is the average expression level of gene i over all p experiments and $\parallel D_i \parallel = \sqrt{\sum_{e=1}^{p}(D(i,e) - \mu_i)^2}$ is the norm of the expression vector D_i with the mean subtracted. Correlation coefficients range from -1 to 1. Two genes with correlation coefficient equals to 1 are perfectly correlated, *i.e.,* their expression levels change in the same direction across the experiments. On the other hand, a correlation coefficient of -1 means that two genes are anti-correlated, *i.e.,* their expression levels change in opposite directions. Geometrically, correlation coefficients capture the *patterns* of expression levels of two genes. For example, two genes with different average expression levels but with expression levels peaking at the same experiments have a high correlation coefficient.

The Euclidean distance between a pair of genes i and j $(1 \leq i, j \leq n)$ is defined as

$$\sqrt{\sum_{e=1}^{p}(D(i,e) - D(j,e))^2} \tag{16.2}$$

A high Euclidean distance between a pair of genes indicates low similarity between the genes. Unlike correlation coefficients, Euclidean distances measure both the direction and amplituide difference in expression levels. For example, two genes peaking at the same experiments but with different average expression levels may lead to a large Euclidean distance, especially if the difference in average expression levels is high.

4. Clustering algorithms

There is a rich literature in clustering algorithms, and there are many different classifications of clustering algorithms. One classification is *model-based* versus *heuristic-based* clustering algorithms. The objects to be clustered are assumed to be generated from an underlying probability framework in the model-based clustering approach. In the heuristic-based approach, an underlying probability framework is not assumed. The inputs to a heuristic-based clustering algorithm usually includes the similarity matrix and the number of clusters. CAST and PAM are examples of the heuristic-based approach. The k-means algorithm was originally proposed as a heuristic-based clustering algorithm. However, it was shown to be closely related to the model-based approach (Celeux and Govaert, 1992).

4.1 CAST

The *Cluster Affinity Search Technique (CAST)* (Ben-Dor and Yakhini, 1999), (Ben-Dor et al., 1999) is a graph-theoretic algorithm developed to cluster gene expression data. In graph-theoretic clustering algorithms, the objects to be clustered (genes in this case) are represented as nodes, and pairwise similarities of genes are represented as weighted edges in a graph. The inputs to CAST include the similarity matrix Sim, and a threshold parameter t (which is a real number between 0 and 1), which indirectly controls the number of clusters.

4.1.1 Algorithm outline. CAST is an iterative algorithm in which clusters are constructed one at a time. The current cluster under construction is called C_{open}. The *affinity* of a gene g, $a(g)$, is defined as the sum of similarity values between g and all the genes in C_{open}, *i.e.*, $a(g) = \sum_{x \in C_{open}} Sim(g, x)$. A gene g is said to have high affinity if $a(g) \geq t|C_{open}|$. Otherwise, g is said to have low affinity. Note that the affinity of a gene depends on the genes that are already in C_{open}. When a new cluster C_{open} is started, the initial affinity is zero because C_{open} is empty. A gene not yet assigned to any clusters and having the maximum average similarity to all unassigned genes is chosen to be the first gene in C_{open}. The algorithm alternates between adding high affinity genes to C_{open}, and removing low affinity genes from C_{open}. C_{open} is *closed* when no more genes can be added to or removed from it. Once a cluster is closed, a new C_{open} is formed. The algorithm iterates until all the genes have been assigned to clusters and the current C_{open} is closed. After the CAST algorithm converges (assuming it does), there is an additional iterative step, in which all clusters are considered at the same time, and genes are moved to the cluster with the highest average similarity. For details of CAST, please refer to (Ben-Dor and Yakhini, 1999).

4.1.2 Algorithm properties. Correlation coefficient is usually used as the similarity metric for CAST. From our experience, the iterative step in CAST may not converge if Euclidean distance is used as the similarity metric.

In contrast to the hierarchical clustering approach in which objects are successively merged into clusters, objects can be added to or removed from the current open cluster through the iterative steps. CAST tends to produce relatively high quality clusters, compared to the hierarchical approach (Yeung et al., 2001b).

4.2 K-means

K-means is another popular clustering algorithm in gene expression analysis. For example, Tavazoie *et al.* (Tavazoie et al., 1999) applied k-means to cluster the yeast cell cycle data (Cho et al., 1998).

4.2.1 Algorithm outline. K-means (MacQueen, 1965) is a classic iterative clustering algorithm, in which the number of clusters, k, together with the similarity matrix are inputs to the algorithm. In the k-means clustering algorithm, clusters are represented by *centroids*, which are cluster centers. The goal of k-means is to minimize the sum of distances from each object to its corresponding centroid. In each iteration, each gene is assigned to the centroid (and hence cluster) with the minimum distance (or equivalently maximum similarity). After the gene re-assignment, new centroids of the k clusters are computed. The steps of assigning genes to centroids and computing new centroids are repeated until no genes are moved between clusters (and centroids are not changed). K-means was shown to converge for any metric (Selim and Ismail, 1984).

4.2.2 Effect of initialization. Initialization plays an important role in the k-means algorithm. In the random initialization approach, the k initial centroids consist of k randomly chosen genes. An alternative approach is to use clusters from another clustering algorithm as initial clusters, for example, from hierarchical average-link. The advantage of the second approach is that the algorithm becomes deterministic (the algorithm always yields the same clusters). (Yeung et al., 2001b) showed that the iterative k-means step after the hierarchical step tends to improve cluster quality.

4.2.3 Algorithm properties. Clusters obtained from the k-means algorithm tend to be equal-sized and spherical in shape. This is because the k-means algorithm is closely related to the equal volume spherical model in the model-based clustering approach (Celeux and Govaert, 1992).

4.2.4 Implementation. K-means is implemented in many statistical software packages, including the commercial software Splus (Everitt, 1994), and the GNU free software R (Ihaka and Gentleman, 1996). It is also available from other clustering packages tailored toward gene expression analysis, such as XCluster from Gavin Sherlock, which is available at `http://genome-www.stanford.edu/~sherlock/cluster.html`.

4.3 PAM

Partitioning around Medoids (PAM) (Kaufman and Rousseeuw, 1990) searches for a representative object for each cluster from the data set. These representative objects are called *medoids*. The clusters are obtained by assigning each data point to the nearest medoid. The objective is to minimize the total dissimilarity of objects to their nearest medoid. This is very similar to the objective of k-means, in which the total dissimilarity of objects to their centroids is minimized. However, unlike centroids, medoids do not represent the mean vector of data points in clusters.

4.3.1 Algorithm outline. The inputs to PAM include the similarity or dissimilarity matrix and the number of clusters k. The algorithm of PAM consists of two stages. In the first Build stage, an initial clustering is obtained by successive selection of representative objects until k objects are found. In the second Swap stage, all pairs of objects (i, h), for which object i is in the current set of medoids and object h is not, are considered. The effect on the object function is studied if object h is chosen as a medoid instead of object i.

4.3.2 Algorithm properties. PAM can be considered as a robust version of k-means since medoids are less affected by outliers. Similar to k-means, PAM also tends to produce spherical clusters (Kaufman and Rousseeuw, 1990).

4.3.3 Implementation. PAM is implemented in statistical packages such as Splus and R.

5. Assessment of Cluster Quality

We have discussed three different heuristic-based clustering algorithms to analyze gene expression data. Different clustering algorithms can potentially generate different clusters on the same data set. However, no clustering method has emerged as the method of choice in the gene expression community. A biologist with a gene expression data set is faced with the problem of choosing an appropriate clustering algorithm for his or her data set. Hence, assessing and comparing the quality of clustering results is crucial.

Jain and Dubes (Jain and Dubes, 1988) classified cluster validation procedures into two main categories: external and internal criterion analysis. *External criterion* analysis validates a clustering result by comparing to a given "gold standard" which is another partition of the objects. *Internal criterion* analysis uses information from within the given data set to represent the goodness of fit between the input data set and the clustering results.

5.1 External Validation

In external validation, a clustering result with a high degree of agreement to the "gold standard" is considered to contain high quality clusters. The gold standard must be obtained by an independent process based on information other than the given data. This approach has the strong benefit of providing an independent, hopefully unbiased assessment of cluster quality. On the other hand, external criterion analysis has the strong disadvantage that an external gold standard is rarely available.

Both clustering results and the external criteria can be considered as partitions of objects into groups. There are many statistical measures that assess the agreement between two partitions, for example, the adjusted Rand index (Hubert and Arabie, 1985). The adjusted Rand index is used to assess cluster quality in (Yeung and Ruzzo, 2001) and (Yeung et al., 2001a).

5.2 Internal Validation

Internal criterion analysis does not require an independent external criteria. Instead, it assesses the goodness of fit between the input data set and the clustering results. We will briefly describe three internal validation approaches.

5.2.1 Homogeneity and separation. Since objects in the same cluster are expected to be more similar to each other than objects in different groups and objects in different clusters are expected to be dissimilar, *homogeneity* of objects in the same cluster and *separation* between different clusters are intuitive measures of cluster quality (Shamir and Sharan, 2001). Homogeneity is defined as the average similarity between objects and their cluster centers, while separation is defined as the weighted average similarity between cluster centers. A high homogeneity indicates that objects in clusters are similar to each other. A low separation means that different clusters are not well-separated.

5.2.2 Silhouette. *Silhouettes* can be used to evaluate the quality of a clustering result. Silhouettes are defined for each object (gene in our context) and are based on the ratio between the distances of an object to its own cluster and to its neighbor cluster (Rousseeuw, 1987). A high silhouette value indicates that an object lies well within its assigned cluster, while a low silhouette value means that the object should be assigned to another cluster. Silhouettes can also be used to visually display clustering results. The objects in each cluster can be displayed in decreasing order of the silhouette values such that a cluster with many objects with high silhouette values is a pronounced cluster. Silhouettes are implemented in Splus and R. In order to summarize the silhouette values in a data set with k clusters, the *average silhouette width*, is defined to be the average silhouette value over all the objects in the data. The average silhouette width can be used as an internal validation measure to compare the quality of clustering results.

5.2.3 Figure of merit. (Yeung et al., 2001b) proposed the *figure of merit* (FOM) approach to compare the quality of clustering results. The idea is to apply a clustering algorithm to all but one experiment in a given data set, and use the left-out experiment to assess the predictive power of the clustering algorithm.

Intuitively, a clustering result has possible biological significance if genes in the same cluster tend to have similar expression levels in additional experiments that were not used to form the clusters. We estimate this predictive power by removing one experiment from the data set, clustering genes based on the remaining data, and then measuring the within-cluster similarity of expression values in the left-out experiment. The *figure of merit* (FOM) is a scalar quantity, which is an estimate of the predictive power of a clustering algorithm.

6. Model-based approach

Clustering algorithms based on probability models offer a principled alternative to heuristic algorithms. The issues of selecting a “good” clustering method and determining the “correct” number of clusters are reduced to model selection problems in the probability framework. This provides a great advantage over heuristic clustering algorithms, for which there is no rigorous method to determine the number of clusters or the best clustering method. (Yeung et al., 2001a) applied the model-based approach to various gene expression and synthetic data, and showed that the model-based approach tends to produce higher cluster quality than the heuristic-based algorithms.

6.1 The model-based framework

In *model-based* clustering, the data is assumed to be generated from a finite *mixture* of underlying probability distributions[1]. In other words, we assume the data consists of different *groups* (or *components*), and each group (or component) is generated from a known probability distribution. Based on this assumption, the goal of model-based clustering algorithms is to recover clusters that correspond to the components in the data.

There are many possible probability distributions underlying each group (or component). In this chapter, we assume a *Gaussian mixture model* in which each component is generated by the multivariate normal distribution (also known as the multivariate Gaussian distribution)[2]. Gaussian mixture models have been shown to be a powerful tool for clustering in many applications, for example, (Banfield and Raftery, 1993), (Celeux and Govaert, 1993), (McLachlan and Basford, 1988), (McLachlan and Peel, 2000).

The multivariate normal distribution is parameterized by the mean vector μ and covariance matrix Σ. When the objects to be clustered are the genes, the mean vector μ is of dimension p (which is the number of experiments). The mean vector of a component is equal to the average expression level of all the genes in that component. Hence, the mean vector represents the location where the component is centered at. The covariance matrix Σ is a p by p matrix such that $\Sigma(i, j)$ represents the covariance of experiment i and experiment j. The diagonal entries in the covariance matrix are the variances of the p experiments[3].

Let G be the number of components in the data. In the Gaussian mixture assumption, each component k (where $k = 1, \ldots, G$) is generated by the multivariate normal distribution with parameters μ_k (mean vector) and Σ_k (covariance matrix). The number of components G is assumed to be known. The goal is to estimate the parameters μ_k and Σ_k from the data (where $k = 1, \ldots, G$), and find clusters corresponding to these parameter estimates.

In order to make estimation of the parameters easier, (Banfield and Raftery, 1993) proposed a general framework to decompose the covariance matrix

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{16.3}$$

where D_k is an orthogonal matrix, A_k is a diagonal matrix, and λ_k is a scalar. The matrix D_k determines the orientation of the component, A_k determines its shape, and λ_k determines its volume. Hence, the covariance matrix Σ_k controls the shape, volume and orientation of each component.

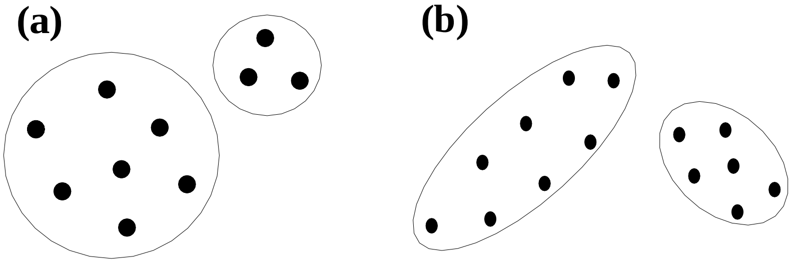

Figure 16.2. Fictitious examples illustrating (a) the unequal volume spherical model in which clusters are spherical but may have different volumes and (b) the unequal volume spherical model in which clusters are spherical but may have different volumes.

Allowing some but not all of the parameters in Equation 16.3 to vary results in a set of models within this general framework. In particular, constraining $D_k A_k D_k^T$ to be the identity matrix I corresponds to Gaussian mixtures in which each component is spherical. For example, the *equal volume spherical* model, which is parameterized by $\Sigma_k = \lambda I$, represents the most constrained model under this framework, with the smallest number of parameters[4]. The classical k-means clustering algorithm has been shown to be closely related to this model (Celeux and Govaert, 1992). However, there are circumstances in which this model may *not* be appropriate. For example, if some groups of genes are much more tightly co-regulated than others, a model in which the spherical components are allowed to have different volumes may be more appropriate. The *unequal volume spherical* model (see Figure 16.2a), $\Sigma_k = \lambda_k I$, allows the spherical components to have different volumes by allowing a different λ_k for each component k. We have also observed considerable correlation between experiments in time-series experiments, coupled with unequal variances. An elliptical model may better fit the data in these cases, for example, the *unconstrained* model (see Figure 16.2b) allows all of D_k, A_k and λ_k to vary between components. The unconstrained model has the advantage that it is the most general model, but has the disadvantage that the maximum number of parameters need to be estimated, requiring relatively more data points in each component. There is a range of elliptical models with other constraints, and hence requiring fewer parameters.

6.2 Algorithm Outline

Assuming the number of clusters, G, is fixed, the model parameters are estimated by the expectation maximization (EM) algorithm. In the

EM algorithm, the expectation (E) steps and maximization (M) steps alternate. In the E-step, the probability of each observation belonging to each cluster is estimated conditionally on the current parameter estimates. In the M-step, the model parameters are estimated given the current group membership probabilities. When the EM algorithm converges, each observation is assigned to the group with the maximum conditional probability. The EM algorithm can be initialized with model-based hierarchical clustering (Dasgupta and Raftery, 1998), (Fraley and Raftery, 1998), in which a maximum-likelihood pair of clusters is chosen for merging in each step.

6.3 Model selection

Each combination of a different specification of the covariance matrices and a different number of clusters corresponds to a separate probability model. Hence, the probabilistic framework of model-based clustering allows the issues of choosing the best clustering algorithm and the correct number of clusters to be reduced simultaneously to a model selection problem. This is important because there is a trade-off between probability model, and number of clusters. For example, if one uses a complex model, a small number of clusters may suffice, whereas if one uses a simple model, one may need a larger number of clusters to fit the data adequately.

Let D be the observed data, and M_k be a model with parameter θ_k. The *Bayesian Information Criterion* (BIC) (Schwarz, 1978) is an approximation to the probability that data D is observed given that the underlying model is M_k, $p(D|M_k)$.

$$2 \log p(D|M_k) \approx 2 \log p(D|\widehat{\theta_k}, M_k) - \nu_k \log(n) = BIC_k \tag{16.4}$$

where ν_k is the number of parameters to be estimated in model M_k, and $\widehat{\theta_k}$ is the maximum likelihood estimate for parameter θ_k. Intuitively, the first term in Equation 16.4, which is the maximized mixture likelihood for the model, rewards a model that fits the data well, and the second term discourages overfitting by penalizing models with more free parameters. A large BIC score indicates strong evidence for the corresponding model. Hence, the BIC score can be used to compare different models.

6.4 Implementation

Typically, different models of the model-based clustering algorithm are applied to a data set over a range of numbers of clusters. The BIC scores for the clustering results are computed for each of the models.

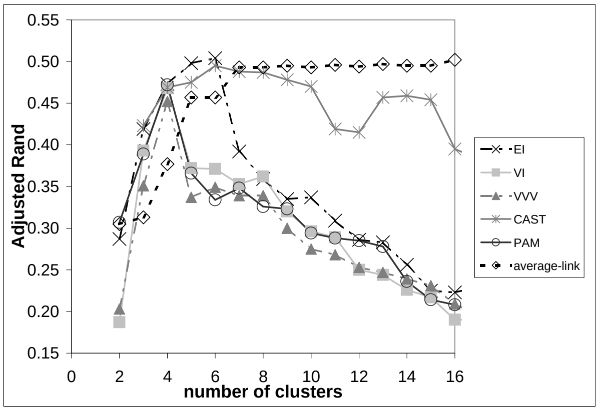

Figure 16.3. Adjusted Rand indices for the standardized yeast cell cycle data.

The model and the number of clusters with the maximum BIC score are usually chosen for the data.

These model-based clustering and model selection algorithms (including various spherical and elliptical models) are implemented in Mclust (Fraley and Raftery, 1998). Mclust is written in Fortran with interfaces to Splus and R. It is publicly available at `http://www.stat.washington.edu/fraley/mclust`.

7. A Case Study

We applied some of the methods described in this chapter to the yeast cell cycle data (Cho et al., 1998), which showed the fluctuation of expression levels of approximately 6000 genes over two cell cycles (17 time points). We used a subset of this data which consists of 384 genes whose expression levels peak at different time points corresponding to the five phases of cell cycle (Cho et al., 1998). We expect clustering results to approximate this five class partition. Hence, the five phases of cell cycle form the external criterion of this data set.

Before any clustering algorithm is applied, the data is pre-processed by standardization, *i.e.,* the expression vectors are standardized to have mean 0 and standard deviation 1 (by subtracting the mean of each row in the data, and then dividing by the standard deviation of the row). Data pre-processing techniques are discussed in detail in Chapter 2.

We applied CAST, PAM, hierarchical average-link and the model-based approach to the standardized yeast cell cycle data to obtain 2 to 16 clusters. The clustering results are evaluated by comparing to the external criterion of the 5 phases of cell cycle, and the adjusted Rand indices are computed. The results are illustrated in Figure 16.3. A high adjusted Rand index means high agreement to the 5-phase external criterion.

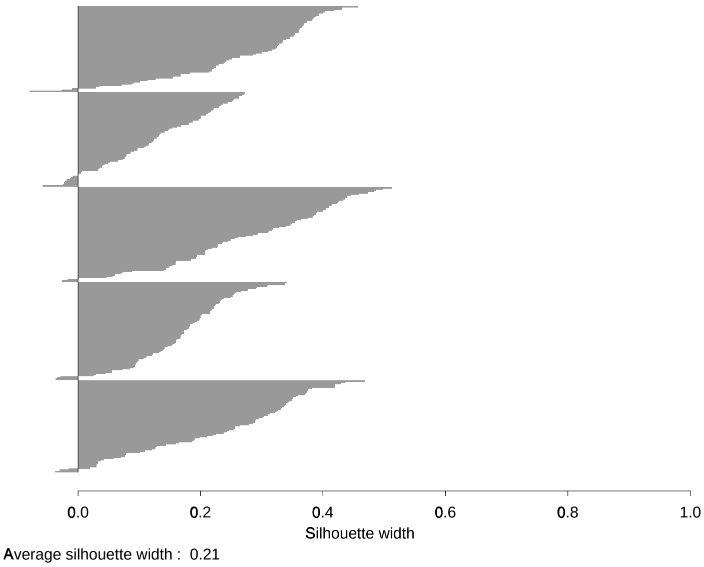

Figure 16.4. A silhouette plot of 5 clusters from PAM on the cell cycle data.

The results from three different models from the model-based approach are shown in Figure 16.3: the equal volume spherical model (denoted by EI), the unequal volume spherical model (denoted by VI), and the unconstrained model (denoted by VVV). The equal volume spherical model (EI) and CAST achieved the highest adjusted Rand indices at 5 clusters. Figure 16.4 shows a silhouette plot of the 5 clusters produced using PAM. Three of the five clusters show higher silhouette values than the other two, and hence, they are relatively more pronounced clusters. In each cluster, there are a few genes with very low silhoutte values, and they represent outliers in the clusters.

Notes

1. A probability distribution is a mathematical function which describes the probability of possible events.

2. A multivariate normal distribution is a generalization of the normal distribution to more than one variables.

3. The variance of an experiment is the average of the squared deviation of the experiment from its mean, while the covariance of two experiments measures their tendency to vary together.

4. Only the parameter λ needs to be estimated to specify the covariance matrix for the equal volume spherical model.

References

Anderberg, M. R. (1973). *Cluster analysis for applications.* Academic Press.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, 6:281–297.

Ben-Dor, A. and Yakhini, Z. (1999). Clustering gene expression patterns. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 33–42, Lyon, France.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332.

Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation*, 47:127–146.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93:294–302.

Everitt, B. (1994). *A handbook of statistical analyses using S-plus.* Chapman and Hall, London.

Everitt, B. S. (1993). *Clustering Analysis.* John Wiley and Sons.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588.

Hartigan, J. A. (1975). *Clustering Algorithms.* John Wiley and Sons.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, New York.

Lander, E. S. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. International Human Genome Sequencing Consortium.

MacQueen, J. (1965). Some methods for classication and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

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

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Selim, S. Z. and Ismail, M. A. (1984). K-means type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1):81–86.

Shamir, R. and Sharan, R. (2001). Algorithmic approaches to clustering gene expression data. In *Current Topics in Computational Biology*. MIT Press.

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

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

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001b). Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318.

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774.