BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1090–1099
DOI: 10.1093/bioinformatics/btg038

Bagging to improve the accuracy of a clustering procedure

Sandrine Dudoit [1,] and Jane Fridlyand [2]*

[1]Division of Biostatistics, School of Public Health, University of California, Berkeley, 140 Earl Warren Hall, #7360, Berkeley, CA 94720-7360, USA and [2]Jain Lab, Comprehensive Cancer Center, University of California, San Francisco, 2340 Sutter St., #N412 , San Francisco, CA 94143-0128, USA

Received on November 15, 2001; revised on November 8, 2002; accepted on November 11, 2002

ABSTRACT

Motivation: The microarray technology is increasingly being applied in biological and medical research to address a wide range of problems such as the classification of tumors. An important statistical question associated with tumor classification is the identification of new tumor classes using gene expression profiles. Essential aspects of this clustering problem include identifying accurate partitions of the tumor samples into clusters and assessing the confidence of cluster assignments for individual samples.

Results: Two new resampling methods, inspired from bagging in prediction, are proposed to improve and assess the accuracy of a given clustering procedure. In these ensemble methods, a partitioning clustering procedure is applied to bootstrap learning sets and the resulting multiple partitions are combined by voting or the creation of a new dissimilarity matrix. As in prediction, the motivation behind bagging is to reduce variability in the partitioning results via averaging. The performances of the new and existing methods were compared using simulated data and gene expression data from two recently published cancer microarray studies. The bagged clustering procedures were in general at least as accurate and often substantially more accurate than a single application of the partitioning clustering procedure. A valuable by-product of bagged clustering are the cluster votes which can be used to assess the confidence of cluster assignments for individual observations.

Contact: sandrine@stat.berkeley.edu

Supplementary information: For supplementary information on datasets, analyses, and software, consult http://www.stat.berkeley.edu/~sandrine and http://www.bioconductor.org.

*To whom correspondence should be addressed.

The authors wish it to be known that, in their opinion, both authors should be regarded as joint First Authors.

1 INTRODUCTION

1.1 Motivation

The burgeoning field of genomics, and in particular DNA microarray experiments, have revived interest in cluster analysis by raising new methodological and computational challenges. Microarray experiments are increasingly being performed in biological and medical research to address a wide range of problems such as the classification of tumors (Alizadeh *et al.*, 2000; Alon *et al.*, 1999; Golub *et al.*, 1999; Perou *et al.*, 1999; Pollack *et al.*, 1999; Ross *et al.*, 2000). An important statistical question associated with tumor classification is the identification of new tumor classes using gene expression profiles. Essential aspects of this clustering problem are: (I) to accurately estimate the number of clusters, if any, in a dataset; (II) to accurately allocate tumor samples to these clusters and assess the confidence of cluster assignments for individual samples. For example, in the context of tumor classification, the definition of new tumor classes is based on the clustering results and these classes could then be used to build predictors for new tumor samples. Inaccurate cluster assignments could lead to errors in diagnosis and assignment to unsuitable treatment protocols.

Resampling methods such as bagging (Breiman, 1996) and boosting (Breiman, 1998; Freund and Schapire, 1997) have been applied successfully in the area of supervised learning to improve prediction accuracy. This and a related article (Dudoit and Fridlyand, 2002) propose resampling procedures to address cluster analysis problems (I) and (II). For question (I), Dudoit and Fridlyand (2002) introduced a novel prediction-based resampling method, *Clest*, to estimate the number of clusters, if any, in a dataset. In the present paper, two bagged clustering procedures are proposed to improve and assess the accuracy of a partitioning clustering method (question (II)). In this context, bagging is used to generate and aggregate multiple clusterings and to assess the confidence of cluster assignments for individual observations. As in prediction,

the motivation behind the application of bagging to cluster analysis is to reduce variability in the partitioning results via averaging. Partitioning methods are typically based on iterative optimization techniques, thus, additional sources of variability in the results include the sensitivity to starting conditions and the possibility of convergence to local minima (or maxima, depending on the objective function). In a recent manuscript, Leisch (1999) proposed a bagged clustering method which is a combination of partitioning and hierarchical procedures. A partitioning method is applied to bootstrap learning sets and the resulting partitions are combined by performing hierarchical clustering of the cluster centers. This procedure compared favorably to existing partitioning methods for a variety of simulated and real datasets considered by the author. Our two new bagging procedures are similar in spirit to that of Leisch (1999), however, different approaches based on voting and the creation of a new dissimilarity matrix are proposed to combine multiple partitioning results.

The article is organized as follows. Section 2 describes two novel resampling methods, *BagClust1* and *BagClust2*, for improving the accuracy of a partitioning clustering procedure and for assessing the confidence of cluster assignments for individual observations. The performances of the proposed bagged clustering procedures and existing methods are compared using simulated data and gene expression data from two recently published cancer microarray studies. The simulation models and microarray datasets are described in Section 3 and the results are given in Section 4. Finally, Section 5 summarizes our findings and outlines open questions. The remainder of this section provides an introduction to partitioning clustering methods and, in particular, to the Partitioning Around Medoids (PAM) clustering method of Kaufman and Rousseeuw (1990) used in the comparison study. Although this article focuses on the clustering of tumors based on gene expression data using *PAM*, the *BagClust1* and *BagClust2* methods are applicable to general clustering problems and procedures (e.g. *k*-means, self-organizing maps).

1.2 Partitioning clustering methods

The data are assumed to be sampled from a mixture distribution with K components corresponding to the K *clusters* to be recovered. Let $(X_1, \ldots, X_p)$ denote the $1 \times p$ random vector of *explanatory variables*, or *features*, and let $Y \in \{1, \ldots, K\}$ denote the unknown component, or *cluster label*. Given a sample of X's, the goal is to estimate the number of clusters K and, for each observation, its cluster label Y. Suppose we have data $\mathbf{X} = (x_{ij})$ on p explanatory variables (e.g. genes) for n observations (e.g. tumor mRNA samples), where x_{ij} denotes the realization of variable X_j for observation i and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ denotes the feature vector for observation i, $i = 1, \ldots, n$, $j = 1, \ldots, p$.

We consider clustering procedures that partition the *learning set* $\mathcal{L} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ into K clusters of observations that are 'similar' to each other, where K is a user prespecified integer. Specifically, a clustering procedure $\mathcal{P}(\cdot; \mathcal{L})$ assigns cluster labels $\mathcal{P}(\mathbf{x}_i; \mathcal{L}) = \hat{y}_i$ to each observation, where $\hat{y}_i \in \{1, \ldots, K\}$. Many clustering algorithms (e.g. PAM) operate directly on a matrix of pairwise *dissimilarities* (or *similarities*) between the observations to be clustered, such as the Euclidean or Manhattan distance matrices (Mardia *et al.*, 1979).

The proposed bagged clustering procedures are illustrated using the *Partitioning Around Medoids* or PAM method of Kaufman and Rousseeuw (1990). As implemented in the `R` and `S-Plus` libraries `cluster`, the two main arguments of the `PAM` function are: a dissimilarity matrix (e.g. the Euclidean distance matrix as used here) and the number of clusters K. The PAM procedure is based on the search for K representative objects, or *medoids*, such that the sum of the dissimilarities of the observations to their closest medoid is minimized. In addition, PAM returns for each observation a *silhouette width* which reflects how well the particular observation is clustered.

2 METHODS

For a given number of clusters K, the goal is to estimate for each observation its cluster label and, if possible, get a measure of confidence for this cluster assignment. In prediction, it is well known that gains in accuracy can be obtained by aggregating predictors built from perturbed versions of the learning set (Breiman, 1996, 1998; Freund and Schapire, 1997). In the *b*ootstrap *agg*regat*ing* or *bagging* procedure (Breiman, 1996), perturbed learning sets of the same size as the original learning set are formed by drawing at random with replacement from the learning set, i.e. by forming non-parametric *bootstrap* replicates of the learning set. Predictors are built for each perturbed dataset and aggregated by plurality *voting*. A useful by-product of the voting are the *prediction votes*, which may be used to assess the confidence of predictions for individual observations (Dudoit *et al.*, 2002). It is of interest to see whether bagging can also lead to increases in accuracy in the context of cluster analysis.

Two bagged clustering procedures, denoted by *BagClust1* and *BagClust2*, are proposed here. In the first method, the clustering procedure is repeatedly applied to each bootstrap sample and the final partition is obtained by *plurality voting*, i.e. by taking the majority cluster label for each observation. Valuable by-products of this bootstrap procedure are the *cluster votes* for individual observations. The second bagging approach forms a *new dissimilarity matrix* by recording for each pair of

observations the proportion of time they were clustered together in the bootstrap clusters (Breiman, pers. comm.). This new dissimilarity matrix is then used as an input to a clustering procedure and the resulting partition is considered final. Note that *BagClust1* and *BagClust2* can be applied to arbitrary clustering procedures; the partitioning clustering procedure PAM is used in this article for illustration purposes.

2.1 Bagged clustering procedure, *BagClust1*

For a fixed number of clusters K

(1) Apply the partitioning clustering procedure $\mathcal{P}$ to the original learning set $\mathcal{L}$ to obtain cluster labels $\mathcal{P}(\mathbf{x}_i; \mathcal{L}) = \hat{y}_i$ for each observation $\mathbf{x}_i$, $i = 1, \ldots, n$.

(2) Form the bth bootstrap sample $\mathcal{L}^b = (\mathbf{x}_1^b, \ldots, \mathbf{x}_n^b)$.

(3) Apply the clustering procedure $\mathcal{P}$ to the bootstrap learning set $\mathcal{L}^b$ and obtain cluster labels $\mathcal{P}(\mathbf{x}_i^b; \mathcal{L}^b)$ for each observation in $\mathcal{L}^b$.

(4) Permute the cluster labels assigned to the bootstrap learning set $\mathcal{L}^b$ so that there is maximum overlap with the original clustering of these observations. Specifically, let S_K denote the set of all permutations of the integers $1, \ldots, K$. Find the permutation $\tau^b \in S_K$ that maximizes
$$\sum_{i=1}^n I\Big(\tau(\mathcal{P}(\mathbf{x}_i^b; \mathcal{L}^b)) = \mathcal{P}(\mathbf{x}_i^b; \mathcal{L})\Big),$$
where $I(\cdot)$ is the indicator function, equaling 1 if the condition in parentheses is true and 0 otherwise.

(5) Repeat Steps 2–4 B times (here, we used $B = 20$) and assign a bagged cluster label for each observation i by *majority vote*, that is, the cluster label corresponding to $\mathbf{x}_i$ is $\text{argmax}_{1\le k\le K} \sum_{\{b:\ \mathbf{x}_i \in \mathcal{L}^b\}} I\Big(\tau^b(\mathcal{P}(\mathbf{x}_i; \mathcal{L}^b)) = k\Big)$. Also, record a *cluster vote*, which is the proportion of votes in favor of the *winning* cluster assignment, that is,
$$CV(\mathbf{x}_i) = \frac{\max_{1\le k\le K} \sum_{\{b:\ \mathbf{x}_i \in \mathcal{L}^b\}} I\Big(\tau^b(\mathcal{P}(\mathbf{x}_i; \mathcal{L}^b)) = k\Big)}{|\{b: \mathbf{x}_i \in \mathcal{L}^b\}|}.$$

The method described next bypasses the alignment in Step 4 by considering pairs of observations, rather than individual observations, and by building a new dissimilarity matrix.

2.2 Bagged clustering procedure, *BagClust2*

For a fixed number of clusters K

(1) Initialize two $n \times n$ matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{M} = (m_{ij})$ to zero.

(2) Form the bth bootstrap sample $\mathcal{L}^b = (\mathbf{x}_1^b, \ldots, \mathbf{x}_n^b)$.

(3) Apply the partitioning clustering procedure $\mathcal{P}$ to the bootstrap learning set $\mathcal{L}^b$ and obtain cluster labels $\mathcal{P}(\mathbf{x}_i^b; \mathcal{L}^b)$ for each observation in $\mathcal{L}^b$.

(4) For each pair of observations, update the matrices $\mathbf{A}$ and $\mathbf{M}$ as follows
$$a_{ij} \leftarrow a_{ij} + I\Big(\mathbf{x}_i \in \mathcal{L}^b,\ \mathbf{x}_j \in \mathcal{L}^b,\ \mathcal{P}(\mathbf{x}_i; \mathcal{L}^b) = \mathcal{P}(\mathbf{x}_j; \mathcal{L}^b)\Big),$$
$$m_{ij} \leftarrow m_{ij} + I\Big(\mathbf{x}_i \in \mathcal{L}^b,\ \mathbf{x}_j \in \mathcal{L}^b\Big).$$

(5) Repeat Steps 2–4 B times (here, we used $B = 20$) and define a new dissimilarity matrix $\mathbf{D} = (d_{ij})$, by $d_{ij} = 1 - a_{ij}/m_{ij}$.

(6) Cluster the n original observations on the basis of this new dissimilarity matrix.

Note that the clustering procedure applied in Step 6 need not be the same as the procedure applied in Step 3. Also, note that unlike *BagClust1*, the *BagClust2* procedure does not directly produce cluster votes. It is nonetheless possible to assess the confidence of the bagged cluster assignments using, for example, the silhouette widths in PAM.

3 IMPLEMENTATION

3.1 Simulation models

The cluster bagging procedures *BagClust1* and *BagClust2* were applied to simulated data and compared, in terms of the accuracy of cluster assignments, to a single application of the clustering procedure PAM.

Observations for each cluster were generated independently from multivariate normal distributions. That is, for each cluster k, n_k independent observations were generated from $\text{N}(\mu_k, \Sigma_k)$, where μ_k and Σ_k denote respectively the $1 \times p$ mean vector and $p \times p$ covariance matrix for cluster k, $k = 1, \ldots, K$. The parameters of the models were set in such a way that the clusters are overlapping to a certain degree. Eight classes of models (referred to as **Models I–VIII**), with varying numbers of variables, covariance matrix structures, and numbers of clusters were considered and listed in Table 1.

Fifty datasets were simulated for each model. For each dataset, three sets of cluster labels were obtained by applying PAM alone (with Euclidean distance metric) and the *BagClust1* and *BagClust2* bagging procedures using PAM and $B = 20$ bootstrap samples. The three partitions were compared to the true partition as follows. The assigned cluster labels of the observations were permuted in order to minimize the proportion of observations with cluster labels disagreeing with the true class labels (see Step 4 of the procedure *BagClust1* in Section 2). The resulting disagreement rate is referred to as the *clustering*

Table 1. Description of simulation models

Model	Cluster mean vectors	Cluster covariance matrices	Cluster sizes	Parameter Δ
Model I				
$K = 2$	$\mu_1 = (0, 0)$	$\Sigma = \mathbf{I}_2$	$n_1 = 50$	1, 3, 6
$p = 2$	$\mu_2 = (0, \Delta)$		$n_2 = 50$	
Model II				
$K = 2$	$\mu_1 = (0, \mathbf{0}_{99})$	$\Sigma = \mathbf{I}_{100}$	$n_1 = 50$	3, 6
$p = 100$	$\mu_2 = (\Delta, \mathbf{0}_{99})$		$n_2 = 50$	
Model III				
$K = 3$	$\mu_1 = \mathbf{0}_{13}$		$n_1 = 50$	1.5, 2
$p = 13$	$\mu_2 = (\Delta, -\Delta, \Delta, \mathbf{0}_{10})$	$\Sigma = \begin{pmatrix} \mathbf{A}_3 & \mathbf{0}_{3,10} \\ \mathbf{0}_{10,3} & \mathbf{I}_{10} \end{pmatrix}$	$n_2 = 50$	
	$\mu_3 = -\mu_2$		$n_3 = 50$	
Model IV				
$K = 3$	$\mu_1 = \mathbf{0}_{13}$		$n_1 = 50$	2
$p = 13$	$\mu_2 = (\Delta, -\Delta, \Delta, \mathbf{0}_{10})$	$\Sigma = \mathbf{A}_{13}$	$n_2 = 50$	
	$\mu_3 = -\mu_2$		$n_3 = 50$	
Model V				
$K = 3$	$\mu_1 = \mathbf{0}_{16}$		$n_1 = 50$	2
$p = 16$	$\mu_2 = (\Delta, 0, \Delta, 0, \Delta, 0, \mathbf{0}_{10})$	$\Sigma = \begin{pmatrix} \mathbf{A}_6 & \mathbf{0}_{6,10} \\ \mathbf{0}_{10,6} & \mathbf{I}_{10} \end{pmatrix}$	$n_2 = 50$	
	$\mu_3 = (0, -\Delta, 0, -\Delta, 0, -\Delta, \mathbf{0}_{10})$		$n_3 = 25$	
Model VI				
$K = 3$	$\mu_1 = \mathbf{0}_{15}$	$\Sigma = \begin{pmatrix} \mathbf{B}_5 & \mathbf{0}_{5,10} \\ \mathbf{0}_{10,5} & \mathbf{I}_{10} \end{pmatrix}$	$n_1 = 50$	1.5, 2
$p = 15$	$\mu_2 = (\Delta, \Delta, \Delta, \Delta, \Delta, \mathbf{0}_{10})$		$n_2 = 50$	
	$\mu_3 = -\mu_2$		$n_3 = 50$	
Model VII				
$K = 3$	$\mu_1 = \mathbf{0}_{15}$	$\Sigma = \begin{pmatrix} \mathbf{C} & \mathbf{0}_{5,10} \\ \mathbf{0}_{10,5} & \mathbf{I}_{10} \end{pmatrix}$	$n_1 = 50$	2
$p = 15$	$\mu_2 = (\Delta, \Delta, \Delta, \Delta, \Delta, \mathbf{0}_{10})$		$n_2 = 50$	
	$\mu_3 = -\mu_2$		$n_3 = 50$	
Model VIII				
$K = 2$	$\mu_1 = \mathbf{0}_{15}$	$\Sigma_1 = \begin{pmatrix} \mathbf{C} & \mathbf{0}_{5,10} \\ \mathbf{0}_{10,5} & \mathbf{I}_{10} \end{pmatrix}$	$n_1 = 50$	2
$p = 15$	$\mu_2 = (\Delta, \Delta, \Delta, \Delta, \Delta, \mathbf{0}_{10})$	$\Sigma_2 = \begin{pmatrix} \mathbf{D} & \mathbf{0}_{5,10} \\ \mathbf{0}_{10,5} & \mathbf{I}_{10} \end{pmatrix}$	$n_2 = 50$	

Here, $\mathbf{0}_{m,n}$ is an $m \times n$ matrix of zeros; $\mathbf{A}_p$ is the $p \times p$ matrix such that $a_{ii} = 1$, and $a_{ij} = 0.5$, $i \neq j$; $\mathbf{B}_p$ is the $p \times p$ matrix such that $b_{ii} = 1$, $b_{i\,i+1} = b_{i\,i-1} = 0.5$, and $b_{ij} = 0.1$, $j \neq i-1, i, i+1$; $\mathbf{C} = \begin{pmatrix} 0.5 & 0.5 & -0.1 & -0.1 & -0.1 \\ 0.5 & 1.0 & 0.5 & -0.1 & -0.1 \\ -0.1 & 0.5 & 1.5 & 0.5 & -0.1 \\ -0.1 & -0.1 & 0.5 & 1.0 & 0.5 \\ -0.1 & -0.1 & -0.1 & 0.5 & 0.5 \end{pmatrix}$; and $\mathbf{D} = \begin{pmatrix} 1.0 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 2.0 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1.0 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 2.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 1.0 \end{pmatrix}$.

error rate. For each simulation model, the three methods were compared based on the distribution of the error rates over the 50 realizations.

For the *BagClust1* procedure we also investigated how well the cluster votes relate to the accuracy of individual cluster assignments. The observations were grouped according to the correctness of their assigned labels and the distributions of the cluster votes for each of the two groups were compared.

3.2 Microarray data

The proposed bagged clustering procedures were applied to gene expression data from two recently published cancer microarray studies (see Table 2): the leukemia (ALL/AML) dataset of Golub *et al.* (1999) and the melanoma dataset of Bittner *et al.* (2000). These and two other datasets are described in detail in Dudoit and Fridlyand (2001) and Dudoit and Fridlyand (2002). The $p = 100$ most variable genes were used for clustering.

Table 2. Description of microarray datasets

Dataset	Number of classes	Class sizes	Number of genes
Leukemia Golub *et al.* (1999) (Affymetrix chips)	$K = 3$ classes	ALL B-cell (38) ALL T-cell (9) AML (25)	$p = 3,571$
Melanoma* Bittner *et al.* (2000) (cDNA microarrays)	$K = 2$ classes	Group A (19) Group B (12)	$p = 3,613$

∗Note that in the leukemia dataset, tumor classes were known *a priori*, while for the melanoma dataset, the two classes were inferred by Bittner *et al.* (2000) via cluster analysis but not confirmed on an independent dataset.

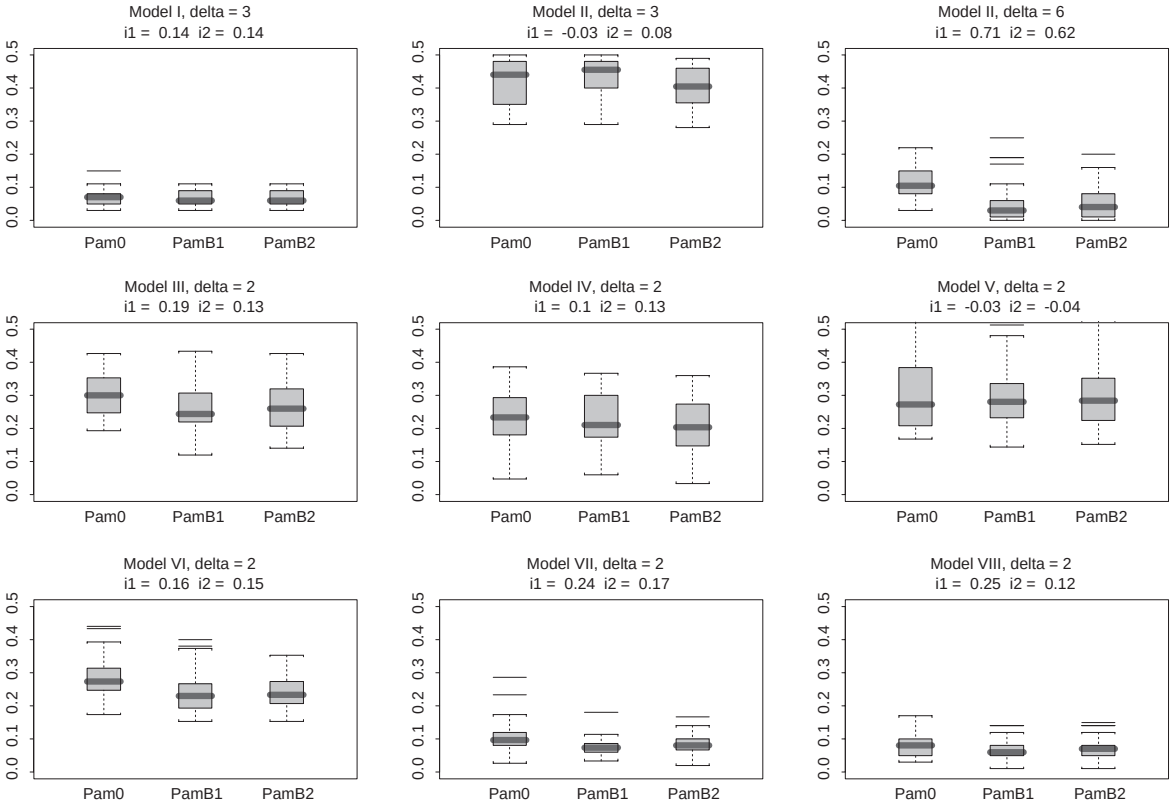

Fig. 1. Improving clustering accuracy, results for simulated data—Boxplots of the clustering error rates (over 50 simulations) for a single application of PAM (Pam0) and bagging procedures *BagClust1* (PamB1) and *BagClust2* (PamB2). Clustering error rates and improvement statistics i_1 and i_2 are defined in Sections 3 and 4.

4 RESULTS

4.1 Comparison of clustering procedures on simulated data

The *BagClust1* and *BagClust2* procedures were evaluated using data simulated from the models described in Section 3.1. The PAM procedure, implemented in the `R` library `cluster`, was used to cluster observations based on the Euclidean distance metric.

4.1.1 Improving clustering accuracy For each simulation model, Figure 1 displays boxplots of the clustering error rates computed over 50 simulations. The clusterings produced by *BagClust1* and *BagClust2* were in general at least as accurate as and often substantially more accurate than the clusterings resulting from a single application of PAM. *BagClust1* and *BagClust2* had very similar performances.

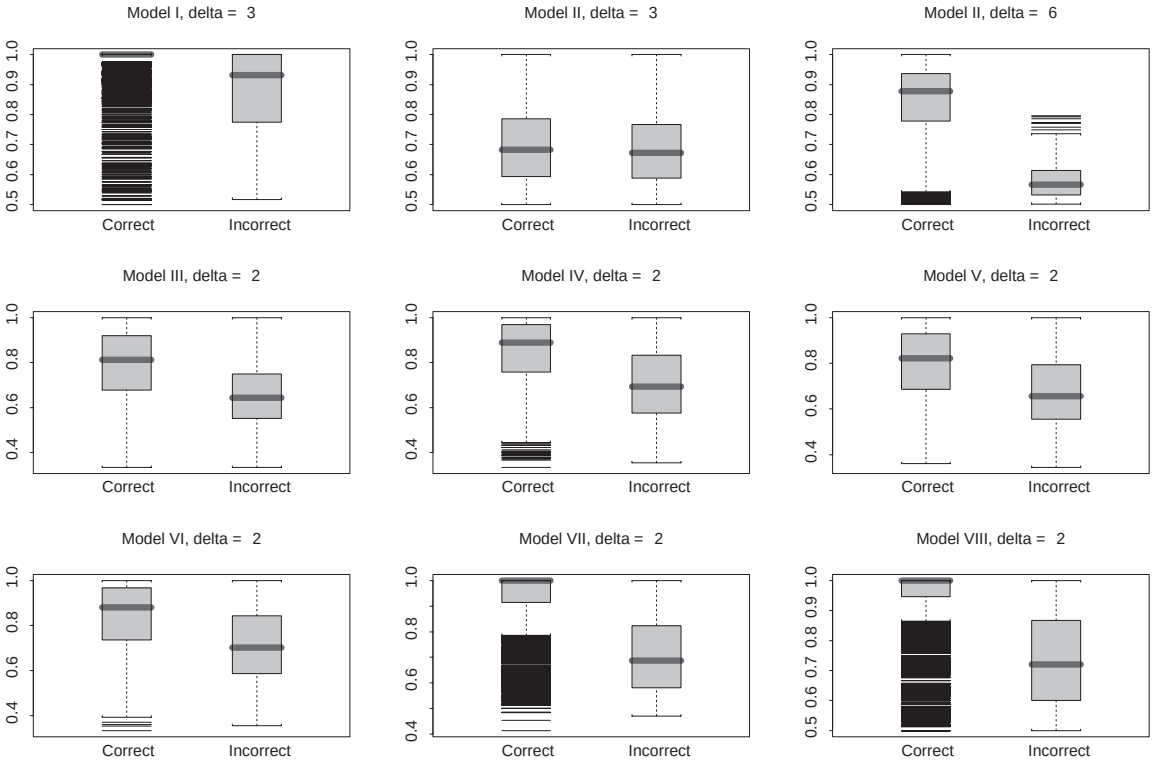

Fig. 2. Cluster votes, results for simulated data—Boxplots of cluster votes for *BagClust1* stratified according to correct and incorrect allocations. (The number of cluster votes considered for each model is equal to the number of observations *n* times the number of simulations, here 50.)

To quantify the improvement of bagging over a single application of PAM, improvement statistics i_1 and i_2 were defined as the percentage change of the clustering error rate relative to a single application of PAM. That is, the *improvement statistic* i_j for *BagClustj*, $j = 1, 2$, is defined as the ratio $(e_o - e_j)/e_0$, where e_0, e_1, and e_2 denote the median clustering error rates for PAM, *BagClust1*, and *BagClust2*, respectively, over the 50 simulated datasets. The improvement statistics are displayed above the boxplots in Figure 1. Bagging resulted in improvements in accuracy of at least 15% for a majority of models, and up to 70% for *BagClust1* applied to **Model II** with $\Delta = 6$. In addition, pairwise comparisons of the clustering error rates for the three procedures were made based on paired t-statistics (data not shown). Overall, *BagClust1* and *BagClust2* had very similar performances: no significant differences were found for five out of the nine models (nominal two-sided z-tests, level 0.05 for each model). For the other four models, the directions of rejection were split evenly in favor of either *BagClust1* or *BagClust2*. *BagClust1* and *BagClust2* were generally superior to PAM alone: significant improvements were seen in 7/9 models for *BagClust1* and 6/9 models for *BagClust2* (one-sided nominal z-tests, level 0.05 for each model).

Both bagging procedures showed the largest improvement over a single application of PAM for **Model II** with $\Delta = 6$ ($i_1 = 71\%$, $i_2 = 62\%$). This model contains a large number of noise variables (99), with complete overlap between the clusters, and only one variable with no overlap between the clusters. The improvement statistics for the bagging procedures were very small and sometimes negative for **Model II** with $\Delta = 3$ and **Model V** with $\Delta = 2$. For these models, aggregation had no impact on the quality of the partitions. In general, the improvement statistic rises as the separation between the clusters increases, unless the performance of a single application of PAM is nearly optimal (data not shown).

4.1.2 Cluster votes Recall that cluster votes CV can be obtained as by-products of the plurality voting in the *BagClust1* procedure. For each model, the observations were stratified according to whether they were correctly clustered by *BagClust1* or not, and the distributions of the cluster votes between the two types of observations were compared using boxplots. From Figure 2, it can be seen that the cluster votes for correctly allocated observations

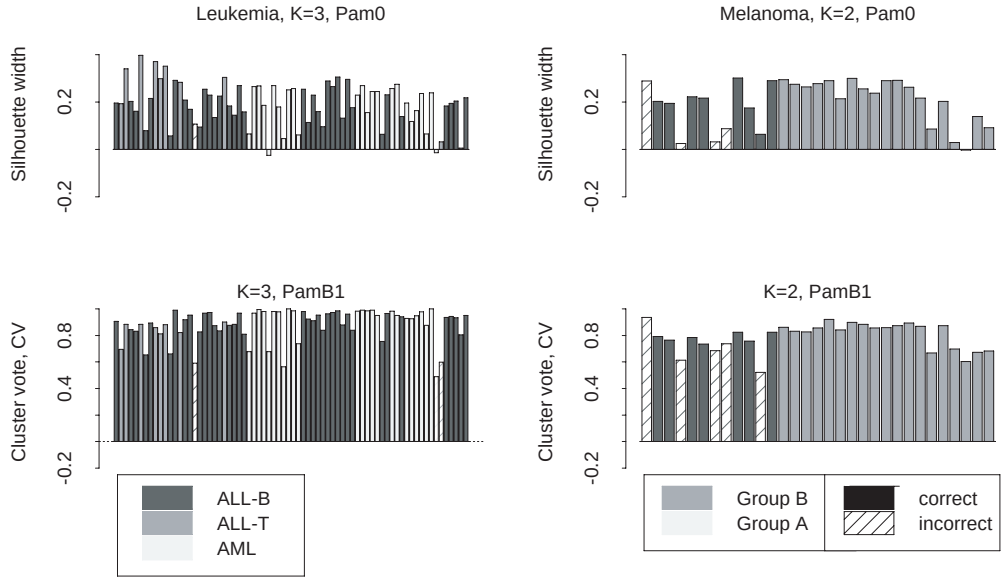

Fig. 3. Cluster votes and silhouette widths, microarray datasets—Silhouette plots for a single application of PAM (Pam0) and plots of cluster votes for *BagClust1* (PamB1). The clusterings are based on the $p = 100$ genes with the largest variances across samples. The 'known' classes of the observations are represented by different colors and incorrect cluster assignments are indicated by hatching. Left-hand column: leukemia dataset; $K = 3$; tumor samples ordered as in Golub *et al.* Right-hand column: melanoma dataset; $K = 2$; tumor samples ordered by class, as proposed in Bittner *et al.* (2000), first Group B, then Group A.

are higher than those for incorrectly allocated ones. Cluster votes are thus good indicators of the accuracy of a cluster assignment.

4.2 Comparison of clustering procedures on microarray data

The *BagClust1* and *BagClust2* procedures were also evaluated using gene expression data from the two cancer microarray studies described in Section 3.2. The PAM clustering procedure of Kaufman and Rousseeuw (1990), implemented in the R library `cluster`, was used to cluster tumor samples based on the Euclidean distance metric. Recall that mRNA samples in the leukemia dataset were assigned class labels from the laboratory analyses of the tumor samples. For the melanoma dataset, tumor class labels were obtained from the statistical analysis described in Bittner *et al.* (2000). In the discussion that follows, these class labels are treated as "truth". In a related article (Dudoit and Fridlyand, 2002), the number of clusters estimated by the *Clest* procedure agreed with the 'known' number of clusters for both datasets. That number was thus used as an input to PAM, *BagClust1*, and *BagClust2*. The resulting cluster assignments and cluster votes are discussed next—a more detailed discussion for these and two other microarray datasets is given in Dudoit and Fridlyand (2001).

4.2.1 Leukemia The PAM, *BagClust1*, and *BagClust2* procedures were applied to the leukemia dataset with $K = 3$ clusters and using the $p = 100$ genes with the largest variances across samples. A single application of PAM clustered one of the AML cases with the ALL T-cell cases; *BagClust1* and *BagClust2* misallocated the same case as Pam0 and clustered one ALL T-cell case with the ALL B-cell cases and one ALL B-cell case with the AML cases. Note that the misallocated cases are the same as the ones that were hard to predict in Dudoit *et al.* (2002) and Golub *et al.* (1999).

Figure 3, left-hand column, displays barplots of the cluster votes and silhouette widths. The silhouette widths tend to be more variable than the cluster votes. Recall that a negative silhouette width indicates that the corresponding observation tends to be closer to observations in clusters other than its own, i.e. its cluster label is suspicious. A few observations that were 'correctly' classified by a single application of PAM have negative or very small silhouette widths; two of these observations were mislabeled by *BagClust1* and carried low cluster votes. This raises the possibility that the tumors were misdiagnosed in the laboratory.

4.2.2 Melanoma The PAM, *BagClust1*, and *BagClust2* procedures were applied to 31 melanoma samples with $K = 2$ clusters and using the $p = 100$ genes

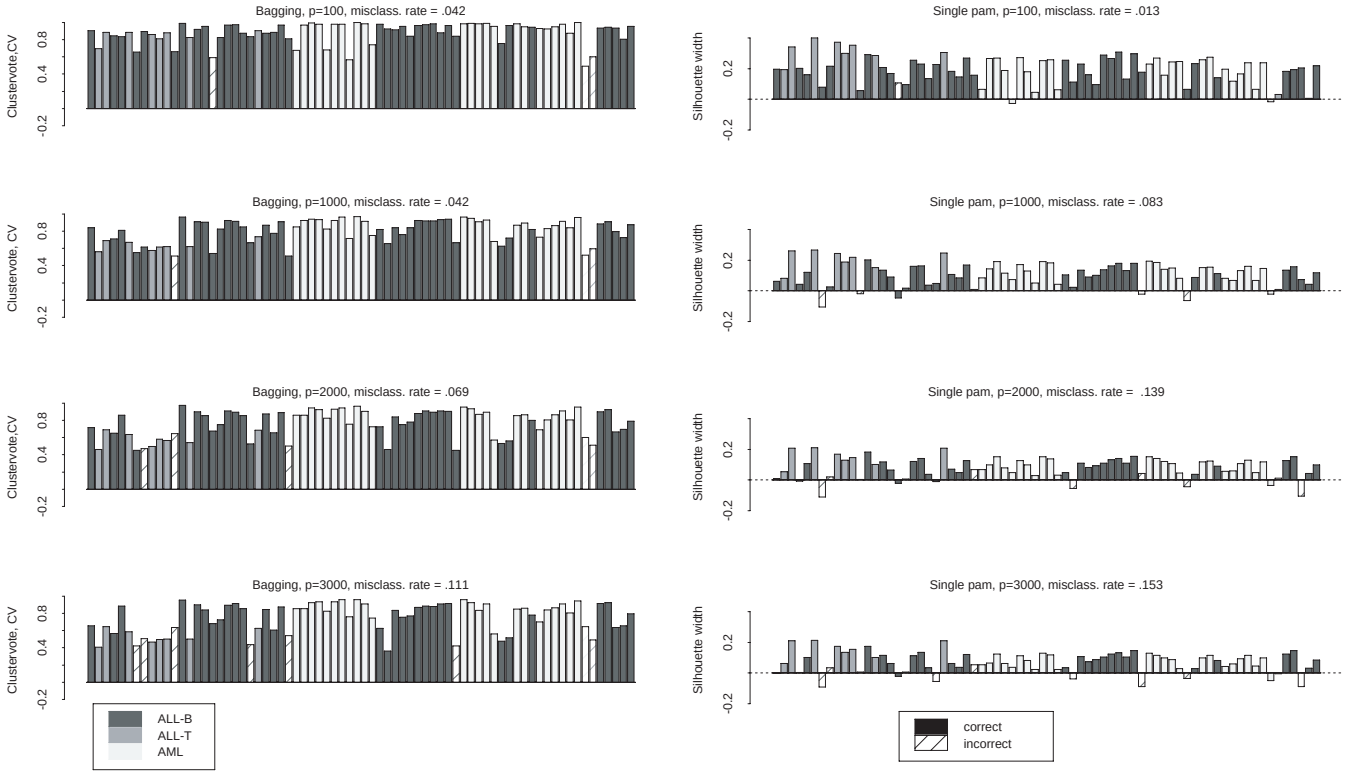

Fig. 4. Cluster votes and silhouette widths, leukemia dataset, $K = 3$, and variable number of genes p—Silhouette plots for a single application of PAM (right-hand column) and plots of cluster votes for *BagClust1* (left-hand column). The clusterings are based on the $p = 100$, 1000, 2000, 3000 genes with the largest variances across samples. Tumor samples are ordered as in Golub *et al.*. Clustering error rates are displayed above each plot.

with the largest variances across samples. Results were compared to the cluster assignments of Bittner *et al.* (2000). Figure 3, right-hand column, displays barplots of the cluster votes and silhouette widths. The *BagClust1* and *BagClust2* partitions were identical. In general, the cluster votes for the melanoma dataset were slightly lower than the cluster votes for the leukemia dataset, providing weaker evidence in favor of the cluster assignments. Four observations allocated by Bittner et al. to the small cluster (Group B) were reclassified to the large cluster (Group A) by PAM, *BagClust1*, and *BagClust2*. For instance, the first observation from the left in Group B was assigned the highest cluster vote and placed into Group A by all three procedures. Bittner et al. reclassified this sample as well in a later analysis (Radmacher, pers. comm.).

To our knowledge, the existence of the two melanoma classes and the correctness of the cluster allocations have not been experimentally or clinically verified. Survival data are available on 15 patients as well as other clinical information. However, these data do not carry enough power to validate the allocation of the melanoma samples into the two clusters.

4.2.3 Effect of number of genes on clustering accuracy In order to study the impact of varying the number of genes on clustering accuracy, the PAM, *BagClust1*, and *BagClust2* procedures were applied to the above datasets using the $p = 100$, 500, 1000, 2000, 3000 genes with the largest variances across tumor samples. In general, the performance of a single application of PAM deteriorated markedly with an increasing number of noise variables (i.e. uninformative genes). In contrast, the *BagClust1* procedure was more robust to the selection of genes and showed a slower decrease in accuracy. Results for the *BagClust2* procedure were intermediate between those of PAM and *BagClust1*, with clustering error rates reaching those of *BagClust1* for a large p. The decrease in accuracy of PAM with an increasing number of genes was also reflected by a decrease in silhouette widths. Results for the leukemia dataset are shown in Figure 4. These results are consistent with findings from the simulation study.

5 DISCUSSION

Resampling methods such as bagging and boosting have been applied successfully in a supervised learning context

to improve prediction accuracy. Here and in a related article (Dudoit and Fridlyand, 2002), we have proposed resampling methods to address two main problems in cluster analysis: (I) estimating the number of clusters, if any, in a dataset; (II) improving and assessing the accuracy of a given clustering procedure. Since the groups obtained from cluster analysis are often used later on for prediction purposes, the approaches to these two problems rely on and extend ideas from supervised learning. Although the methods are applicable to general clustering problems and procedures, particular attention was given to the clustering of tumors using gene expression data. The performances of the proposed and existing procedures were compared using simulated data and gene expression data from four recently published cancer microarray studies (see supplementary information in Dudoit and Fridlyand (2001)).

For problem (II), a resampling method known as *bagging* in supervised learning is used to generate and aggregate multiple clusterings. Two bagged clustering procedures were proposed. In *BagClust1*, the clustering procedure is repeatedly applied to each bootstrap sample and the final partition is obtained by *plurality voting*. The second bagging procedure, *BagClust2*, forms a *new dissimilarity matrix* by recording for each pair of observations the proportion of time they were clustered together in the bootstrap clusters. This new dissimilarity matrix is then used as an input to a clustering procedure (possibly different from the original clustering procedure).

For the microarray and simulated datasets considered in this study, the clusterings produced by bagging procedures *BagClust1* and *BagClust2* were in general at least as accurate and often substantially more accurate than the clusterings resulting from a single application of PAM. *BagClust1* and *BagClust2* were also less sensitive than PAM alone to the number of variables (i.e. genes) used to recover clusters. The greatest improvement in accuracy resulting from bagging occurred in the presence of a large number of noise variables. This suggests that bagged clusterings are less sensitive to the quality of the variable screening procedure. The performances of *BagClust1* and *BagClust2* were very similar.

Although the bagging procedures *BagClust1* and *BagClust2* were illustrated using PAM, they are applicable to any clustering procedure and it would be worthwhile to evaluate the improvement in accuracy for methods such as *k*-means or self-organizing maps. We suspect that, as in prediction, the increase in accuracy observed with PAM is due to a decrease in variability achieved by aggregating multiple clusterings. It would be interesting to carry out a more thorough study of the bias and variance properties of different clustering methods, as was done for classifiers in Breiman (1998). Other ongoing research directions include the investigation of different resampling schemes, similar in spirit to the adaptive resampling schemes used in boosting.

Valuable by-products of the *BagClust1* procedure are the *cluster votes* for individual observations. Our study indicates that cluster votes are generally good indicators of the accuracy of a cluster assignment. In the context of tumor microarray data, samples with low cluster votes could be 'flagged' and sent for new laboratory analyses. The cluster votes could be used as weights when building predictors for the classes obtained by clustering. Note that one could also compute for a given observation the distribution of the cluster votes for each cluster and interpret the results as in *fuzzy clustering* (see Kaufman and Rousseeuw (1990) for a discussion of fuzzy clustering).

An interesting feature of the *BagClust1* procedure was raised in the application to the NCI 60 dataset using $K = 8$ clusters (Dudoit and Fridlyand, 2001). Although each application of PAM to a bootstrap learning set produced eight clusters, the plurality voting reduced the number of clusters to 2. This suggests that *BagClust1* may be able to correct for a misspecified number of clusters by eliminating unstable clusters through the voting step. To investigate this more thoroughly one would need to carry out a simulation study in which the wrong number of clusters is given as an input to *BagClust1*. We are also exploring other methods for combining the bootstrap clustering results in Steps 4 and 5 of the procedure.

Recall that *BagClust2* replaces the original dissimilarity matrix by a bagged dissimilarity matrix. The resulting dissimilarity measure is not a metric; in particular, it fails to satisfy the definiteness property since one can have $d_{i,j} = 0$ for $i \neq j$. While this new dissimilarity measure seems to lead to more accurate clustering results, its theoretical properties (e.g. consistency) remain to be examined. We are further exploring the general idea of creating new dissimilarity matrices by resampling. This could lead to dissimilarity matrices that are more robust to the initial choice of metric and pre-processing decisions such as standardization. The new dissimilarity matrices could be used as inputs to other clustering procedures than the one used on the bootstrap samples.

6 CONCLUSIONS

Application of bagging to cluster analysis can substantially improve clustering accuracy and yields information on the accuracy of cluster assignments for individual observations. In addition, bagged clustering procedures are more robust to the variable selection scheme, i.e. their accuracy is less sensitive to the number and type of variables used in the clustering.

ACKNOWLEDGMENTS

The authors are most grateful to Leo Breiman for many insightful discussions on the topic of classification. They would also like to thank four anonymous referees for their constructive comments on an earlier version of the article. This work was supported in part by a PMMB Burroughs-Wellcome postdoctoral fellowship (SD) and a PMMB Burroughs-Wellcome graduate fellowship (JF).

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. *et al.* (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. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bittner,M., Meltzer,P., Chen,Y., Jiang,Y., Seftor,E., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Breiman,L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman,L. (1998) Arcing classifiers. *Annals of Statistics*, **26**, 801–824.

Dudoit,S. and Fridlyand,J. (2001) Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. *Technical Report 600*. Department of Statistics, Uiversity of California, Berkeley.

Dudoit,S. and Fridlyand,J. (2002) A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biol.*, **3**, 0036.1–0036.21.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Freund,Y. and Schapire,R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.

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

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

Leisch,F. (1999) Bagged clustering. *Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science*. Vienna University of Economics and Business Administration, http://www.ci.tuwien.ac.at/~leisch/papers/fl-techrep.html

Mardia,K.V., Kent,J.T., Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, San Diego.

Perou,C.M., Jeffrey,S.S., van de Rijn,M., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C.F. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **9**, 9212–9217.

Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D., Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Spellman,P., Iyer,V., Jeffrey,S.S., de Rijn,M.V., Waltham,M., Pergamenschikov,A. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–234.