Biostatistics (2008), **9**, 1, *pp.* 128–136
doi:10.1093/biostatistics/kxm017
Advance Access publication on June 14, 2007

Microarray learning with ABC

DHAMMIKA AMARATUNGA*

Johnson & Johnson Pharmaceutical Research & Development LLC, Raritan, NJ 08869-0602, USA
damaratu@prdus.jnj.com

JAVIER CABRERA, VLADIMIR KOVTUN

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA

SUMMARY

Standard clustering algorithms when applied to DNA microarray data often tend to produce erroneous clusters. A major contributor to this divergence is the feature characteristic of microarray data sets that the number of predictors (genes) in such data far exceeds the number of samples by many orders of magnitude, with only a small percentage of predictors being truly informative with regards to the clustering while the rest merely add noise. An additional complication is that the predictors exhibit an unknown complex correlational configuration embedded in a small subspace of the entire predictor space. Under these conditions, standard clustering algorithms fail to find the true clusters even when applied in tandem with some sort of gene filtering or dimension reduction to reduce the number of predictors. We propose, as an alternative, a novel method for unsupervised classification of DNA microarray data. The method, which is based on the idea of aggregating results obtained from an ensemble of randomly resampled data (where both samples and genes are resampled), introduces a way of tilting the procedure so that the ensemble includes minimal representation from less important areas of the gene predictor space. The method produces a measure of dissimilarity between each pair of samples that can be used in conjunction with (a) a method like Ward's procedure to generate a cluster analysis and (b) multidimensional scaling to generate useful visualizations of the data. We call the dissimilarity measures ABC dissimilarities since they are obtained by aggregating bundles of clusters. An extensive comparison of several clustering methods using actual DNA microarray data convincingly demonstrates that classification using ABC dissimilarities offers significantly superior performance.

Keywords: Classification; Clustering; Random forest; Weighted random sampling.

1. INTRODUCTION

Classification techniques, both supervised and unsupervised, have been among the most popular and useful tools for extracting information from the complex mega-variate data sets characteristic of modern high throughput functional genomics research technologies like DNA microarrays and protein arrays. Even the seminal paper on DNA microarrays by Eisen *and others* (1998) involved classifying a set of microarray

*To whom correspondence should be addressed.

samples. Specific details are provided in recent microarray data analysis texts such as Amaratunga and Cabrera (2004) and Gentleman *and others* (2005).

Despite the ready availability of a large variety of classification techniques in the multivariate data analysis and machine-learning domains (Hastie *and others* (2001) and Seber (1984) are good references), the somewhat exceptional structure of microarray data often renders standard procedures inefficient. Some particular issues are that (a) the number of predictors (i.e. genes) far exceeds the number of samples, (b) only a small percentage of the available predictors are truly predictive, and (c) the predictors exhibit an unknown complex correlational configuration.

For supervised classification, extensive comparison studies conducted by Dudoit *and others* (2002) and Lee *and others* (2005) reported relatively high error rates for classical methods such as the Fisher linear discriminant analysis and classification and regression trees. On the other hand, both studies reported procedures based on aggregating classifiers obtained from ensembles of randomly resampled data, notably random forest (Breiman, 2001), as having generally good performance.

High error rates occur with standard procedures in unsupervised classification or clustering as well, with error rates exceeding 20%, even 40%, as demonstrated in Section 4. Nevertheless, careful analysis with unsupervised classification has proven useful in microarray settings as a means of identifying structures in the data (D'haeseleer (2005) discusses clustering in gene expression work). It has been possible to discern disease subtypes, notably in cancer, such as in cutaneous malignant melanoma (Bittner *and others*, 2000), B-cell lymphoma (Alizadeh *and others*, 2000), brain glioblastoma (Mischel *and others*, 2003), and breast cancer (Kapp *and others*, 2006). Even in experimental situations in which the classes are known beforehand, it is still instructive to perform a cluster analysis as a proof of concept or an experimental quality check to ensure that the gene expression profiles differentiate enough so that the expected groups indeed manifest as separate clusters.

In unsupervised classification, as with supervised classification, explorations as to whether aggregating results obtained from ensembles of randomly resampled data would result in improved classifications have proven positive. In these, Dudoit and Fridlyand (2003) use bagging (Breiman, 1996) and Breiman and Cutler (2003), Shi and Horvath (2006), and Shi *and others* (2005) use random forest on artificially augmented data as a means of generating a set of proximity measures by recording how often each pair of samples cluster together; these proximity measures are subsequently used as similarities (or dissimilarities) in a clustering algorithm based on similarities (or dissimilarities).

A key aspect of the problem is feature selection. Biological intuition affirms that of all the genes arrayed on a microarray, only a small subset is truly relevant for differentiating among the different classes of samples. Eliminating the others, which merely obscure the picture by contributing noise, should vastly improve classification accuracy. Recognizing this, gene filtering is often performed, for instance, by filtering out genes with low variance or coefficient of variation (e.g. using Bioconductor's genefilter library), but the results tend to vary markedly with the number of genes selected and fail to take into account combinations of genes that could be useful. It is desirable to have an approach that is less influenced by gene selection, considers combinations of genes in some way, and is automatic and more adaptive to the data.

This then is the motivation for proposing a novel method for unsupervised classification (or learning) that builds up on the idea of aggregating results obtained from an ensemble of randomly resampled data (but doing so without the need for data augmentation). We introduce a technique of tilting the ensemble so that it includes minimal representation from less important regions of the gene predictor space, essentially a form of feature selection but one that is less drastic than the commonly used tack of gene filtering. The resampling involves both samples and genes, with a subset of genes being drawn at each iteration weighted toward genes with higher variance. Each resampled subset is used for a cluster analysis, and the resulting bundle of clusters is used, as above, to generate dissimilarity measures that are subsequently used as dissimilarities in a clustering algorithm based on dissimilarities. We call the dissimilarity measures ABC dissimilarities as they are based on Aggregating Bundles of Clusters.

We demonstrate, using actual DNA microarray data, that classification using ABC dissimilarities offers significantly superior performance.

2. THE PROCEDURE

As is customary in the DNA microarray literature, we assume that the data have been organized as a gene expression matrix, $X=\{x_{gi}\}$, whose G rows and N columns represent, respectively, G genes and N samples. Depending on the experiment, the N samples may correspond to N tissues, cell lines, patients, or other biological samples. The entries x_{gi} are the measured gene expression levels for the gth gene in the ith sample, suitably transformed and normalized (following the methods of Amaratunga and Cabrera (2004) for preprocessing). The objective is to cluster the N samples.

The procedure involves generating a bundle of clusters as follows: Let $n = \sqrt{N}$ and $g = \sqrt{G}$. The procedure is iterative. At the rth iteration:

1) Calculate a weight w_k for each gene k: $w_k = 1/(R_k + c)$, where R_k is the rank of the variance of the kth gene and c is such that the 1% of genes with the highest variance have a combined probability of 20% of being selected.
2) Select N samples with replacement, discarding any replicate samples. Let $J^*_{ijr} = 1$ if the ith and jth samples of X are both selected, $J^*_{ijr} = 0$ otherwise. Let $\sum_{r=1}^{R} J^*_{ijr} = N^*$.
3) Select g genes using weighted random sampling without replacement with weights w_k.
4) Run Ward's clustering procedure on the resulting $g \times N^*$ matrix X^* to find n clusters. These are the rth "base clusters."
5) Record $I^*_{ijr} = 1$ if the ith and jth samples of X cluster together, $I^*_{ijr} = 0$ otherwise.

Repeat steps (1)–(5) R times (say $R = 1000$) and let $P_{ij} = \sum_{r=1}^{R} I^*_{ijr} \big/ \sum_{r=1}^{R} J^*_{ijr}$. A relatively small value of P_{ij} would indicate that samples i and j are relatively close to each other, while a relatively large value of P_{ij} would indicate that samples i and j are relatively far from each other. Thus, $D_{ij} = 1 - P_{ij}$ can be used as a dissimilarity measure, which we refer to as the ABC-dissimilarity measure, and as an input to a hierarchical clustering algorithm, which in our implementation is Ward's method (Seber, 1984). Further details of the procedure are available at *Biostatistics* online.

3. PROPERTIES

A crucial aspect of the procedure is the use of weights calculated in step (1) and used in step (3). The randomness of the subsets used for building the base clusters increases the diversity of the assemblage and reduces dependence among base clusters. However, a large majority of genes in a typical microarray data set are non-informative, and if simple random sampling were used, most of the gene subsets selected will consist mostly of non-informative genes, potentially depleting the accuracy of the base clusters. In a microarray experiment in which only H of G genes contain information about the desirable grouping, if at each iteration we select g genes by resampling with equal weights, the probability distribution of the number of informative genes selected is binomial with g trials and probability $\pi = H/G$ so that the mean number of informative genes selected at each iteration is $\mu = \pi g$. Since π is typically small, so will μ be. For example, if $H = 100$ and $G = 10\,000$ ($g = G^{1/2} = 100$), μ is only one gene per iteration. By weighting the random sampling of genes so that less informative genes are less likely to get selected, our procedure tilts the odds in favor of selecting informative genes and thus is able to include a high number of such genes at all iterations and thereby form better base clusters. The enormous choice of predictors available ensures that the diversity of the ensemble (necessary for the effectiveness of ensemble methods;

Breiman, 1996, 1997) is not compromised. The value of doing the weighting is amply demonstrated in the performance assessment reported in Section 3.

Dissimilarity measures used in clustering need not be metrics (Sibson, 1971). For example, "1-correlation," and the random forest dissimilarity are not. Neither is the ABC dissimilarity, d_{ABC}. Nevertheless, it satisfies the following properties: $0 \leqslant d_{\mathrm{ABC}}(x_1, x_2) \leqslant 1$, $d_{\mathrm{ABC}}(x_1, x_1) = 0$, $d_{\mathrm{ABC}}(x_1, x_2) = d_{\mathrm{ABC}}(x_2, x_1)$. In addition, when run without bagging (i.e. step 2 involving the resampling of arrays), d_{ABC} also satisfies the triangle inequality: $d_{\mathrm{ABC}}(x_1, x_2) \leqslant d_{\mathrm{ABC}}(x_1, x_3) + d_{\mathrm{ABC}}(x_3, x_2)$ (see *Biostatistics* online for proof).

4. Performance assessment

A number of data sets for which the sample groupings are known were used to compare the performance of our procedure with several others. The data sets we used were chosen because they are reasonably representative of common microarray experiments, are fairly different in size and group numbers, and (with one exception) are available publicly. They are as follows:

Golub: The Golub data set, which is one of the earliest and the most widely used data sets in the microarray literature, comes from an experiment in which gene expression levels were measured for 3 types of acute leukemia tumor samples: 38 B-cell acute lymphoblastic leukemia (B-ALL) samples, 9 T-cell acute lymphoblastic leukemia (T-ALL) samples, and 25 acute myeloid leukemia (AML) samples, using Affymetrix GeneChips containing 7129 human genes (Golub *and others*, 1999). The data were preprocessed, cleaned, and scaled before analysis leaving 5888 working genes.

AMS: This data set comes from an experiment in which gene expression levels were measured for 3 types of leukemia tumor samples: 24 acute lymphoblastic leukemia (ALL) samples, 20 AML samples, and 28 mixed lineage leukemia (MLL) samples, using Affymetrix GeneChips containing 8700 human genes (Armstrong *and others*, 2001). All the 8700 genes were included in the analysis.

ALL: The ALL data set consists of data for 128 patients with recently diagnosed ALL: 95 B-cell patients and 33 T-cell patients (Chiaretti *and others*, 2004). The analysis is run on all the 12 625 genes.

Colon: The Colon data set is a colon cancer data set consisting of 2000 genes and 62 samples. The 62 samples came from 40 normal and 22 cancerous colon tissues. Affymetrix GeneChips were used in the experiment (Alon *and others*, 1999).

CL: This data set comes from an experiment in which gene expression levels of 20 samples were measured using Codelink microarrays with 34 946 genes. The 20 samples fall into 5 groups of 4 samples each, the different groups corresponding to different treatments (data provided by T. Shi, unpublished data).

Iris: Finally, to demonstrate the versatility of the ABC approach, we chose a familiar non-microarray data set, one often used in the multivariate statistics literature. This is the famous iris data set (Fisher, 1936) which consists of the measurements of the lengths and widths of the sepals and petals (i.e. 4 predictors) of 150 flowers, which can be classified into 3 species of iris with 50 flowers in each. This data set offers a challenge as 2 of the species are contiguous in the predictor space and are tricky to separate.

Armed with the knowledge of the true nature of each sample, we compared the relative performance of several common clustering algorithms by determining how many samples were misclassified. The methods compared are the following.

The primary procedure and variations:

(1) The method as described above (referred to as BagWeight).

(2) No sampling of arrays and unequal weight sampling of genes (i.e. omit step (2)) (NoBagWeight)

(3) Sampling of arrays and no sampling of genes (i.e. omit steps (1) and (3), essentially the approach taken by Dudoit and Fridlyand (2003)) (BagNoWeight)

(4) no sampling of arrays and equi-weight sampling of genes (i.e. set $w_k = 1$ in step (1) of the procedure) (NoBagNoWeight)

(5) sampling of arrays and equi-weight sampling of genes (BagWholeData).

Other methods (as comparators):

(6)–(16) Several hierarchical clustering methods (using both Euclidean (Euc) and 1-correlation (Cor) as dissimilarities), K-means, partitioning around medoids (PAM) (Kaufman and Rousseeuw, 1990), and random forest clustering.

The results are shown in Table 1. The complete method with both sampling of arrays and weighted sampling of genes (i.e. “BagWeight”) clearly did best: it was the top performer for all but one microarray data set and the second best performer in that. In fact, a major find is that the 2 methods that use weighted sampling (i.e. “BagWeight” and “NoBagWeight”) were among the 2 or 3 best performers across all the microarray data sets and for these data sets generally performed considerably better than any standard method. Of the non-ensemble methods, PAM and Ward’s method gave the best results. The popular hierarchical clustering methods were among the worst performers.

The ABC dissimilarity between 2 samples, in general, quite accurately reflects the dissimilarity between them if their gene expression patterns are reasonably similar. Thus, ABC dissimilarities for pairs

Table 1. *Misclassification rates (expressed as percentages) with the lowest and second lowest rates for each data set displayed in bold. The time column shows the computational time for each method (in s), averaged over the* 6 *data sets. All computations were done in R (version* 2.2.1*). We wrote R routines for methods (*1*)–(5) and (*16*), and since R code tends to run slower than C code by about* 10%*, the time entries for these methods are shown in italics*

Method	Golub	AMS	ALL	Colon	CL	Iris	Time
(1) BagWeight	**18.1**	**1.4**	**0.0**	**9.7**	**20**	9.3	*82.02*
(2) NoBagWeight	**18.1**	**2.8**	**0.0**	21.0	**20**	10.0	*35.00*
(3) BagNoWeight	29.1	6.9	3.9	48.4	**20**	10.7	*40.83*
(4) NoBagNoWeight	25.0	4.2	**1.6**	27.4	**20**	10.0	*32.33*
(5) BagWholeData	**16.7**	4.2	2.3	48.4	**20**	9.3	*303.25*
(6) Single linkage (Cor)	47.0	47.0	25.0	37.0	40	34.0	0.16
(7) Single linkage (Euc)	51.0	58.0	25.0	37.0	40	32.0	0.29
(8) Complete linkage (Cor)	37.5	23.6	41.4	45.0	40	14.7	0.17
(9) Complete linkage (Euc)	25.0	23.6	41.4	45.0	40	16.0	0.22
(10) Average linkage (Cor)	47.2	27.8	26.5	38.7	40	**5.3**	0.16
(11) Average linkage (Euc)	51.3	27.8	26.5	38.7	40	9.3	0.43
(12) Ward’s method (Cor)	23.6	9.7	2.3	48.4	40	**7.3**	0.29
(13) Ward’s method (Euc)	29.2	9.7	40.0	48.4	40	9.3	0.15
(14) K-means	20.8	5.5	42.2	48.4	40	10.7	0.49
(15) PAM	23.6	8.3	2.3	**16.1**	60	10.7	0.20
(16) Random forest	43.0	26.4	48.0	43.5	**20**	10.0	*1227.25*

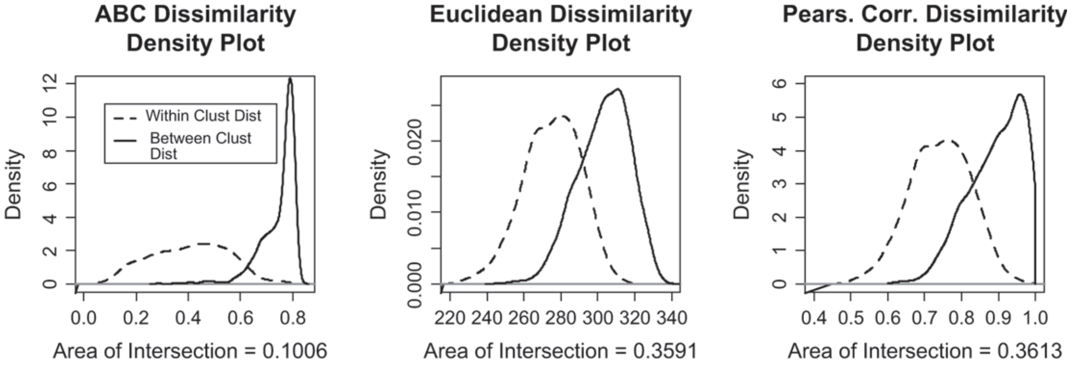

Fig. 1. Density plots of the intracluster and intercluster dissimilarities for (a) ABC dissimilarities, (b) Euclidean distance, and (c) 1-Pearson correlation for the AMS data set. The fraction of the area that lies in the overlap is shown below each plot.

of samples belong to the same class range from being very small for nearby samples (so that they cluster together very often in the base clusters) to being moderately large for samples that are further apart (so that they cluster together a little less often in the base clusters) so that, in such cases, the ABC dissimilarities tend to properly reflect the differences between samples. This is not so when the samples are distant. In this case, pairs of samples belonging to different classes rarely tend to fall into base clusters together so that ABC dissimilarities for pairs of samples belonging to different classes tend to take values close to the maximum value of one. In such cases, the ABC dissimilarity does not provide an accurate measure of how dissimilar the sample profiles are. This tendency is evident in Figure 1, which shows the density plots of the intercluster and intracluster ABC dissimilarities for the AMS data set. It is this tendency and the ensuing small overlap between intercluster and intracluster dissimilarities that make ABC dissimilarities so effective for cluster identification. In contrast, observe that the Euclidean distance and the Pearson correlation (the corresponding plots are also shown in Figure 1) do not exhibit this tendency, have greater overlap between intercluster and intracluster dissimilarities, and are hence less effective.

5. Visualization

For data exploration, it is very useful to be able to visually render and examine the results of a classification procedure. A simple way to do this is to use ABC dissimilarities to generate a configuration of points in 2-dimensional space via multidimensional scaling (see e.g. Kendall, 1971, or Seber, 1984) and then to plot this configuration on a scatter plot. Figure 2 shows such configurations for the Golub and AMS data sets. Except for a few samples, the clusters clearly separate. Contrast this with the usual principal components analysis (PCA) plots for the same data sets shown in Figure 3, where the clusters are almost impossible to discern. For example, for the AMS data set, the ABC plot separates out all the 3 groups, whereas the PCA plot only separates out MLL. (Notes: (a) The addition of a third principal component did not improve the separation. (b) A spectral map (Wouters *and others*, 2003) is an alternative PCA-like display that does show a separation of the clusters.) The tendency of unalike samples to have more or less equal ABC-dissimilarity values (as described above) results in the ABC plots exhibiting a “horseshoe” shape, a term coined by Kendall (1971), who saw such patterns in multidimensional scaling plots with dissimilarities having this property.

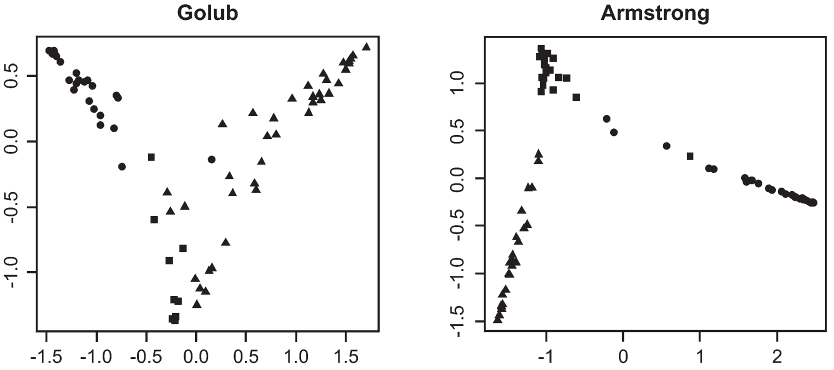

Fig. 2. Multidimensional scaling plots of the ABC dissimilarities for the Golub and AMS data sets; the actual groups are identified by different symbols: for the Golub data set, ●, AML; ▲, B-ALL; and ■, T-ALL; for the AMS data set, ●, MLL; ▲, ALL; and ■, AML.

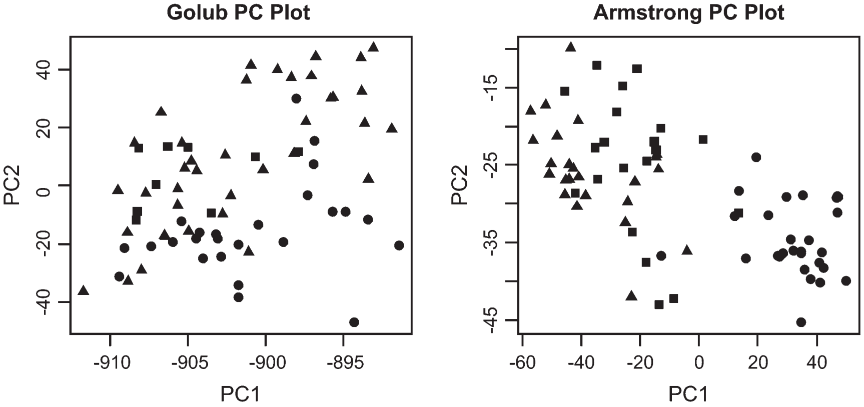

Fig. 3. Principal components plots of the Golub and AMS data sets; the actual groups are identified by different symbols as above.

6. Discussion

We have introduced an ensemble method based on weighted resampling for unsupervised learning. We have applied it to several actual experimental data sets and found it to be highly effective. We conclude with a few miscellaneous comments.

Computational considerations and software: ABC dissimilarities can be computed at relatively low cost even in a data set with a very large number of genes due to the gene selection involved at each iteration. Parallelization is possible as the cluster analyses in step (4) could be run in parallel. An R script (non-parallelized) for computing ABC dissimilarities are available at the authors' Web sites: http://www.rci.rutgers.edu/∼cabrera/DNAMR/, http://www.geocities.com/damaratung/.

As can be seen from Table 1, ABC dissimilarities are very much faster to compute than random forest dissimilarities.

Number of clusters: In settings where the number of clusters is not known, a number of methods may be used in conjunction with ABC dissimilarities to determine the number of clusters, including Dunn's (1974) statistic, the Davies–Bouldin (1979) index, and the silhouette (Rousseeuw, 1987). These methods all gave reasonable, though not perfect, results for the above data sets. Since determining the optimum number of clusters is an entire research topic in itself, this is not studied further here.

Supervised classification with ABC dissimilarities: When the classes are known, ABC dissimilarities can be used in a dissimilarity-based supervised classification scheme for prediction. Initial trials showed that this gave good results. Since it is an entire topic in itself, this will be explored elsewhere.

References

Alizadeh, A., Eisen, A., Davis, M. B., Ma, R. E., Lossos, C., Rosenwald, I. S., Boldrick, A., Sabet, J. C., Tran, H., Yu, T. *and others* (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. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 6745–6750.

Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. New York: John Wiley.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R. and Korsmeyer, S. J. (2001). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics* **30**, 41–47.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. *and others* (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. (1997). Pasting bites together for prediction in large data sets and online. *Technical Report*. Berkeley, CA: Department of Statistics, University of California.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. and Cutler, A. (2003). Random forests manual (version 4.0). *Technical Report*. Berkeley, CA: Department of Statistics, University of California.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771–2778.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions Pattern Analysis and Machine Intelligence* **1**, 224–227.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology* **23**, 1499–1501

Dudoit, S. and Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* **19**, 1090–1099.

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

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* **4**, 95–104.

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

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

GENTLEMAN, R., CAREY, V., HUBER, W., IRIZARRY, R. AND DUDOIT, S. (2005). *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A. *and others* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

KAPP, A. V., JEFFREY, S. S., LANGEROD, A., BORRESEN-DALE, A. L., HAN, W., NOH, D. Y., BUKHOLM, I. R., NICOLAU, M., BROWN, P. O. AND TIBSHIRANI, R. (2006). Discovery and validation of breast cancer subtypes. *BMC Genomics* **7**, 231.

KAUFMAN, L. AND ROUSSEEUW, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons

KENDALL, D. G. (1971). Seriation from abundance matrices. In: Hodson, F. R., Kendall, D. G. and Tautu, P. (editors), *Mathematics in the archaeological and historical sciences: Proceedings of the Anglo-Romanian Conference, Mamaia, 1970*. Edinburgh, Scotland: Edinburgh University Press.

LEE, J. W., LEE, J. B., PARK, M. AND SONG, S. H. (2005). An extensive evaluation of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis* **48**, 869–885.

MISCHEL, P. S., SHAI, R., SHI, T., HORVATH, S., LU, K. V., CHOE, G., SELIGSON, D., KREMEN, T. J., PALOTIE, A., LIAU, L. *and others* (2003). Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene* **217**, 2361–2372.

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.

SEBER, G. A. F. (1984). *Multivariate Observations*. New York: John Wiley & Sons.

SHI, T. AND HORVATH, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics* **15**, 118–138.

SHI, T., SELIGSON, D., BELLDEGRUN, A. S., PALOTIE, A. AND HORVATH, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology* **18**, 547–557.

SIBSON, R. (1971). Some observations on a paper by Lance and Williams. *Computational Journal* **14**, 156–157.

WOUTERS, L., GÖHLMANN, H. W., BIJNENS, L., KASS, S. U., MOLENBERGHS, G. AND LEWI, P. J. (2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* **59**, 1133–1141.

[*Received December 18, 2006; revised April 5, 2007; accepted for publication April 11, 2007*]