Incremental Model-Based Clustering for Large Datasets With Small Clusters

Chris FRALEY, Adrian RAFTERY, and Ron WEHRENS

Clustering is often useful for analyzing and summarizing information within large datasets. Model-based clustering methods have been found to be effective for determining the number of clusters, dealing with outliers, and selecting the best clustering method in datasets that are small to moderate in size. For large datasets, current model-based clustering methods tend to be limited by memory and time requirements and the increasing difficulty of maximum likelihood estimation. They may fit too many clusters in some portions of the data and/or miss clusters containing relatively few observations. We propose an incremental approach for data that can be processed as a whole in memory, which is relatively efficient computationally and has the ability to find small clusters in large datasets. The method starts by drawing a random sample of the data, selecting and fitting a clustering model to the sample, and extending the model to the full dataset by additional EM iterations. New clusters are then added incrementally, initialized with the observations that are poorly fit by the current model. We demonstrate the effectiveness of this method by applying it to simulated data, and to image data where its performance can be assessed visually.

Key Words: BIC; EM algorithm; Image; MRI.

1. INTRODUCTION

The growing size of datasets and databases has led to increased demand for good clustering methods for analysis and compression, while at the same time introducing constraints in terms of memory usage and computation time. Model-based clustering, a relatively recent development (McLachlan and Basford 1988; Banfield and Raftery 1993; McLachlan and Peel 2000; Fraley and Raftery 2002), has shown good performance in many applications on small-to-moderate-sized datasets (e.g., Campbell et al. 1997, 1999; Dasgupta and Raftery 1998; Mukherjee et al. 1998; Yeung et al. 2001; Wang and Raftery 2002; Wehrens, Buydens, Fraley, and Raftery 2004).

Chris Fraley is Senior Research Scientist, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195. Adrian Raftery is Professor of Statistics and Sociology, Box 354322, University of Washington, Seattle, WA 98195. Ron Wehrens is Associate Professor, Department of Analytical Chemistry, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands.

Journal of Computational and Graphical Statistics, Volume 14, Number 3, Pages 529–546
DOI: 10.1198/106186005X59603

Direct application of model-based clustering to large datasets is often prohibitively expensive in terms of computer time and memory. Instead, extensions to large datasets usually rely on modeling one or more random samples of the data, and vary in how the sample-based results are used to derive a model for all of the data. Underfitting (not enough groups to represent the data) and overfitting (too many groups in parts of the data) are common problems, in addition to excessive computational requirements. In this article we develop an incremental model-based method that is suitable as a general clustering method for large datasets that are not too large to be processed as a whole in core (currently up to about 100,000 observations for a dataset of dimension 10 or less), and is also able to find small clusters if they are present.

This article is organized as follows. Section 2 gives a brief overview of model-based clustering and introduces the incremental method. Section 3 gives results for three large datasets. The first is a large simulated dataset with 14 large clusters and 1 small cluster. The remaining two examples are images to be segmented automatically by clustering information associated with each pixel. One of the images has a prominent feature involving only a small number of pixels, making segmentation challenging while at the same time easy to assess visually. The other image is a brain MRI dataset where the task is to find features of anatomical and clinical interest. Section 4 discusses our results and alternative approaches to the problem.

2. METHODS

2.1 MODEL-BASED CLUSTERING

In model-based clustering, the data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are assumed to be generated by a mixture model with density

$$\prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \, f_k(\mathbf{x}_i \mid \theta_k),$$

where $f_k(\mathbf{x}_i \mid \theta_k)$ is a probability distribution with parameters θ_k, and τ_k is the probability of belonging to the kth component or cluster. Most often (and throughout this article) the f_k are taken to be multivariate normal distributions, parameterized by their means μ_k and covariances Σ_k:

$$f_k(\mathbf{x}_i \mid \theta_k) = \phi(\mathbf{x}_i \mid \mu_k, \Sigma_k) = |2\pi\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) \right\},$$

where $\theta_k = (\mu_k, \Sigma_k)$.

The parameters of the model are often estimated by maximum likelihood using the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997). Each EM iteration consists of two steps, an E-step and an M-step.

Given an estimate of the component means μ_j, covariances Σ_j and mixing proportions τ_j, the E-step computes the conditional probability that object i belongs to cluster k:

$$z_{ik} = \phi(\mathbf{x}_i|\mu_k, \Sigma_k) / \sum_{j=1}^{G} \phi(\mathbf{x}_i|\mu_j, \Sigma_j) \ .$$

In the M-step, parameters are estimated from the data given the conditional probabilities z_{ik} (see, e.g., Celeux and Govaert 1995). The E-step and M-step are iterated until convergence, after which an observation can be assigned to the component or cluster corresponding to the highest conditional or posterior probability. Good initial values for EM can be obtained efficiently for small-to-moderate-sized datasets via model-based hierarchical clustering (Banfield and Raftery 1993; Dasgupta and Raftery 1998; Fraley 1998).

Banfield and Raftery (1993) expressed the covariance matrix for the kth component or cluster in the form

$$\mathbf{\Sigma}_k = \lambda_k D_k A_k D_k^T ,$$

where D_k is the matrix of eigenvectors determining the orientation, A_k is a diagonal matrix proportional to the eigenvalues determining the shape, and λ_k is a scalar determining the volume of the cluster. They used this formulation to define a class of hierarchical clustering methods based on cross-cluster geometry, in which mixture components may share a common shape, volume, and/or orientation. This approach subsumes a number of existing clustering methods. For example, if the clusters are restricted to be spherical and identical in volume, the clustering criterion is the same as that which underlies Ward's method (Ward 1963) and k-means clustering (MacQueen 1967). Banfield and Raftery (1993) developed this class of models in the context of hierarchical clustering estimated using the classification likelihood, but the same parameterizations can also be used with the mixture likelihood. A detailed description of the 14 different models that are possible under this scheme can be found in Celeux and Govaert (1995).

Several measures have been proposed for choosing the clustering model (parameterization and number of clusters); see, for example, McLachlan and Peel (2000, chap. 6). The BIC approximation to the Bayes factor (Schwarz 1978), which adds a penalty to the log-likelihood based on the number of parameters, has performed well in a number of applications (e.g., Dasgupta and Raftery 1998; Fraley and Raftery 1998, 2002).

The following strategy for model-based clustering has been found to be effective for datasets of up to moderate size:

Basic Model-Based Clustering Strategy

1. Specify the minimum and maximum number of clusters to consider, $(G_{\text{min}}, G_{\text{max}})$, and a set of candidate parameterizations of the Gaussian model.
2. Do EM for each parameterization and each number of clusters $G_{\text{min}}, \ldots, G_{\text{max}}$, starting with conditional probabilities corresponding to a classification from unconstrained model-based hierarchical clustering.

3. Compute BIC for the mixture likelihood with the optimal parameters from EM for $G_{\min}, \ldots, G_{\max}$ clusters.
4. Select the model (parameterization/number of clusters) for which BIC is maximized.

For a review of model-based clustering, see Fraley and Raftery (2002).

A limitation of the basic model-based clustering strategy for large datasets is that the most efficient computational methods for model-based hierarchical clustering have storage and time requirements that grow at a faster than linear rate relative to the size of the initial partition, which is usually the set of singleton observations. Banfield and Raftery (1993) used hierarchical model-based clustering to cluster a random sample of the data in a large image, and then extended the results to the rest of the data using discriminant analysis. Similarly, the basic model-based clustering strategy can be extended to large datasets in several ways using a random sample of the data. For example, Step 2 of the basic model-based clustering strategy can be modified to do hierarchical clustering only on a random sample of the data, rather than on the whole dataset (Fraley and Raftery 1998, 2002). The corresponding parameter estimates are then used as initial values for EM on the whole dataset. We call this method Strategy **W**.

Strategy W (model from the whole dataset)

1. As in the basic model-based clustering strategy.
2. Do EM for each parameterization and each number of clusters $G_{\min}, \ldots, G_{\max}$, starting with parameters obtained from an M-step with conditional probabilities corresponding to a classification from unconstrained model-based hierarchical clustering on a random sample of the data.

3, 4. As in the basic model-based clustering strategy.

Strategy **W** is of limited practical interest because of its excessive computational requirements (e.g., Wehrens et al. 2004), and we consider it here only for comparison purposes.

Another sample-based alternative is to apply the basic model-based clustering strategy to a random sample of the data to choose the model and number of clusters, and then extend that model via EM to the whole of the data (Fraley and Raftery 1998, 2002). We call this method Strategy **S**.

Strategy S (model from a sample)

1. Apply the basic model-based clustering strategy to a random sample of the data.
2. Extend the result to the whole dataset via EM. [Extension is possible via a single E-step (in which case the whole of the data need not be in core) or one or more EM iterations.]

EM is applied to the full dataset only once for a given initial sample in Strategy **S**, in contrast to Strategy **W**, in which EM is run to convergence on the whole dataset for each model/number of clusters combination.

Wehrens et al. (2004) showed that Strategy **S** can be improved upon by considering several candidate models and several EM steps (rather than just one) in the extension to the full dataset. However, a drawback of Strategy **S** is that it may miss clusters that are small but nevertheless significant.

2.2 INCREMENTAL MODEL-BASED CLUSTERING

In order to improve on the ability to detect small clusters, we propose an incremental procedure which starts with a model that underestimates the number of components (e.g., a model obtained from Strategy **S**), and successively attempts to add new components. EM is initialized with the observations that have the lowest density under the current model in a separate component, and with the rest of the observations initially in their current most probable component, and iterated to convergence. The idea is that the new component would consist largely of observations that are poorly fit by the current mixture model. The process is terminated after an attempt to add a component results in no further improvement in BIC. We call this method Strategy **I**:

Strategy I (incremental model-based clustering)

1. Obtain an initial mixture model for the data that underestimates the number of components (e.g., from Strategy **S**).
2. Choose a set Q of observations with the lowest densities under the current mixture model (e.g., those corresponding to the lowest 1% of densities).
3. Run one or more steps of EM starting with conditional probabilities corresponding to a discrete classification with observations in Q in a separate group, and the rest of the observations grouped according to the current model.
4. If Step 3 results in a higher BIC, go to Step 2. Otherwise fit a less parsimonious model (if possible) starting with the current classifiation and go to Step 2.
5. If the current model is unconstrained, and BIC decreases in Step 4, stop and take the highest-BIC model to be the solution.

The choice of an initial model is required in Step 1. In the examples shown here, we used a model from Strategy **S**, which extends a model-based clustering of a randomly selected sample of the data to the whole dataset via EM. We tried several alternatives for initialization based on random partitions of the data, but the resulting clusterings were not as good as those from initialization with Strategy **S**, and had lower BIC values.

In Step 2, we used the observations with the lowest 1% of densities under the current model to initialize the new component. We experimented with other specifications, both adaptive and nonadaptive, and found that using the lowest 1% worked as well as or better than alternatives.

The model change provision in Step 3 attempts to preserve parsimony in the model as much as possible; in practice the algorithm often terminates with an unconstrained covariance model.

3. RESULTS

Whenever a random sample of the data was required in the algorithms, we used a sample of size 2,000, based on the practical limitation on sample size due to the initial hierarchical clustering, as well as on extensive experiments in both real and simulated datasets for Strategies **S** and **W** and some variants (Wehrens et al. 2004). EM was run to convergence throughout, as defined by a relative tolerance of 10^{-5} in the log-likelihood.

We used the `MCLUST` software (Fraley and Raftery 1999, 2003) for both model-based hierarchical clustering and EM for mixture models. For the image data, only the equal shape `VEV` (varying (`V`) volume, equal (`E`) shape, and varying (`V`) orientation) and unconstrained `VVV` model were considered, because these seemed to be the only models chosen for these data (Wehrens et al. 2004).

For examples we use two-dimensional simulated data, as well as two different real image datasets, to allow visual assessment. Because each method starts by drawing a random sample of observations from the data, we repeated the computations for 100 different initial samples, so that our conclusions do not depend on the particular sample drawn.

3.1 SIMULATED DATA

These data, shown in Figure 1, consist of 50,000 data points generated from a mixture of 14 multivariate normal clusters with equal volume and shape, plus an additional 10-point

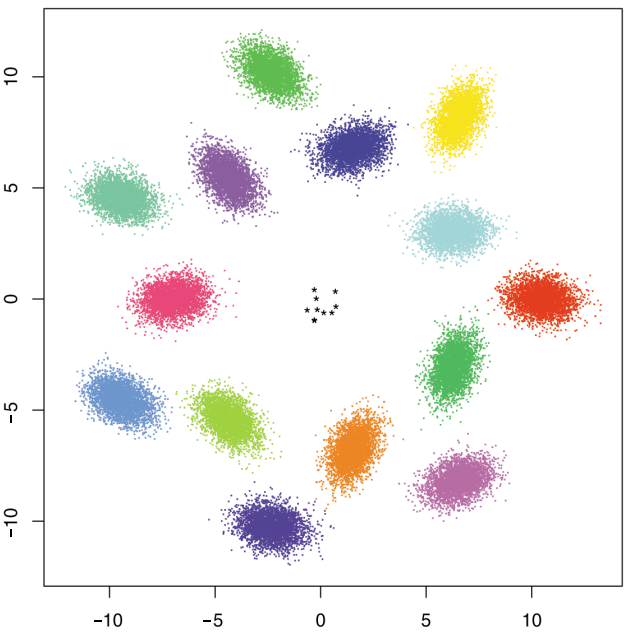


Figure 1. Data simulated from a multivariate normal mixture in which the components have equal volume and shape. There are 50,000 points in 14 clusters away from the center, and 10 points in the cluster near the center. Larger symbols are used for the points near the center to improve visibility. The data consist of spatial coordinates only and do not include color information. Colors are shown here to indicate the true clustering.

Table 1. Results for the Model-Based Clustering Strategies on the Simulated Data. One hundred different random initial samples were used; "sd" denotes the standard deviation. 451,600 was added to all the BIC values for legibility*. Strategy **S** misses the small central cluster for many of the initial samples. Strategy **W** fits more components than are present in the underlying mixture model in some cases. Strategy **I** obtains the global maximum-likelihood estimate for all initial samples.

Strategy	*Mean*	*sd*	*min*	*max*
		Shifted BIC		
S	−514	296	−752	−76
I	−76	0	−76	−76
W	−101	49	−267	−76
		Number of clusters		
S	14.4	.5	14	15
I	15	0	15	15
W	15.3	.7	15	18

* Because BIC is on the logarithmic scale, the main interest is in the differences between values. Adding a constant to all BIC values does not change the conclusions drawn.

cluster centered at the origin, generated from a multivariate normal distribution with the same volume and shape.

Results for Strategies **S**, **I**, and **W** on the simulated data of Figure 1 are shown in Table 1. Strategy **S** chose a model with 14 clusters for 67 out of 100 different random initial samples, missing the small cluster in the middle. In these 14-cluster groupings, each point in the small central cluster is assigned to one of the larger surrounding clusters (Figure 2).

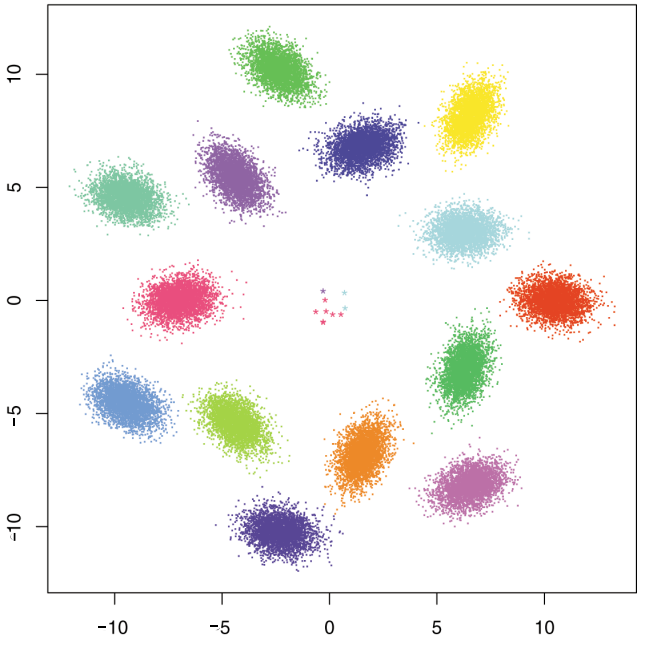


Figure 2. Fourteen-group classification produced by strategy **S**. *Each of the 10 points in the small central cluster is assigned to one of the larger surrounding clusters. Larger symbols are used for the central points to make them visible.*

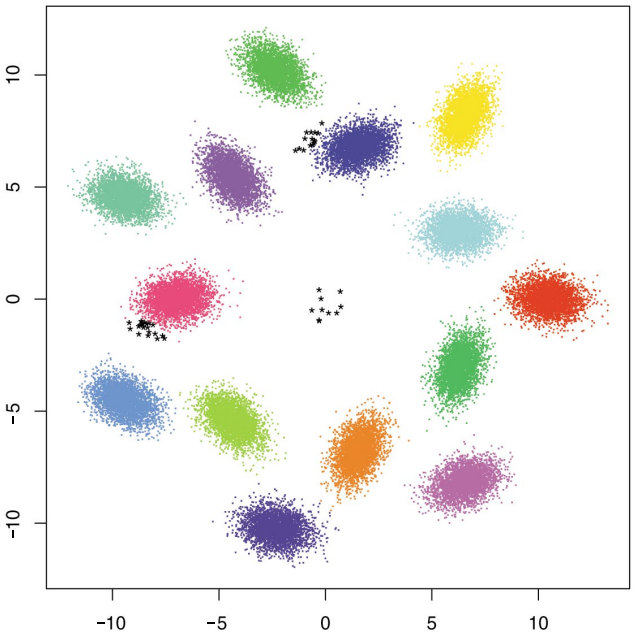


Figure 3. A 17-cluster result from applying strategy **W** *to the simulated data of Figure 1. In addition to the central cluster, there are two other small clusters (also shown in black) away from the center. Larger symbols are used for the three small clusters to improve visibility.*

Strategy **I** classified the data correctly for 100 out of 100 initial samples.

Although it might seem that results from Strategy **W** would be close to the best achievable, it has shortcomings that go beyond its computational requirements. For the simulated data of Figure 1, Strategy **W** did not always select the mixture model for which BIC is maximized (see Table 1). Despite considerable extra computation, Strategy **W** found a model with BIC lower than the maximum in 27 out of 100 cases, due either to local minima in the likelihood surface, or to unusually slow convergence of EM. In 14 of these cases, the resulting classification was virtually indistinguishable from the maximum BIC classification: either the equal shape `VEV` model was chosen instead of the underlying equal-shape equal-volume model, or else the number of components in the mixture was greater than 15, but the conditional probabilities mapped into the correct classification. In 13 other cases, Strategy **W** classified the data into more than 15 clusters; for an example, see Figure 3.

3.2 St. Paulia Flower Image Data

These data describe an RGB (three-band) image of a St. Paulia flower shown in Figure 4. [These data were obtained from J. Noordam, Agrotechnology Innovations B. V., Wageningen, The Netherlands, and are available for downloading from http://www.cac.science.ru.nl/people/rwehrens/suppl/mbc.html.] There are 46,656 pixels after background removal. The small yellow flower centers are particularly eye-catching.

Results for the model-based clustering strategies on the flower image of Figure 4 are

Figure 4. RGB image of a St. Paulia flower. There are 46,656 data pixels after background removal. Note the small yellow flower centers.

Table 2. Results for the Model-Based Clustering Strategies on the St. Paulia Flower Image Data. One hundred different random initial samples were used; "sd" denotes the standard deviation. 1,188,000 was added to all the BIC values. Segmentations produced by Strategy **S** miss the yellow flower centers for all of the initial samples. Segmentations produced by Strategy **W** reveal the yellow centers in only 15 out of 100 initial samples. Segmentations produced by Strategy **I** reveal the yellow centers for 99 out of 100 different samples.

Strategy	*Mean*	*sd*	*min*	*max*
		Shifted BIC		
S	−8577	1646	−14644	−5031
I	−858	344	−2976	−435
W	−568	145	−1053	−317
		Number of clusters		
S	9.6	1.5	7	13
I	22.3	1.9	16	28
W	30.0	3.4	25	39

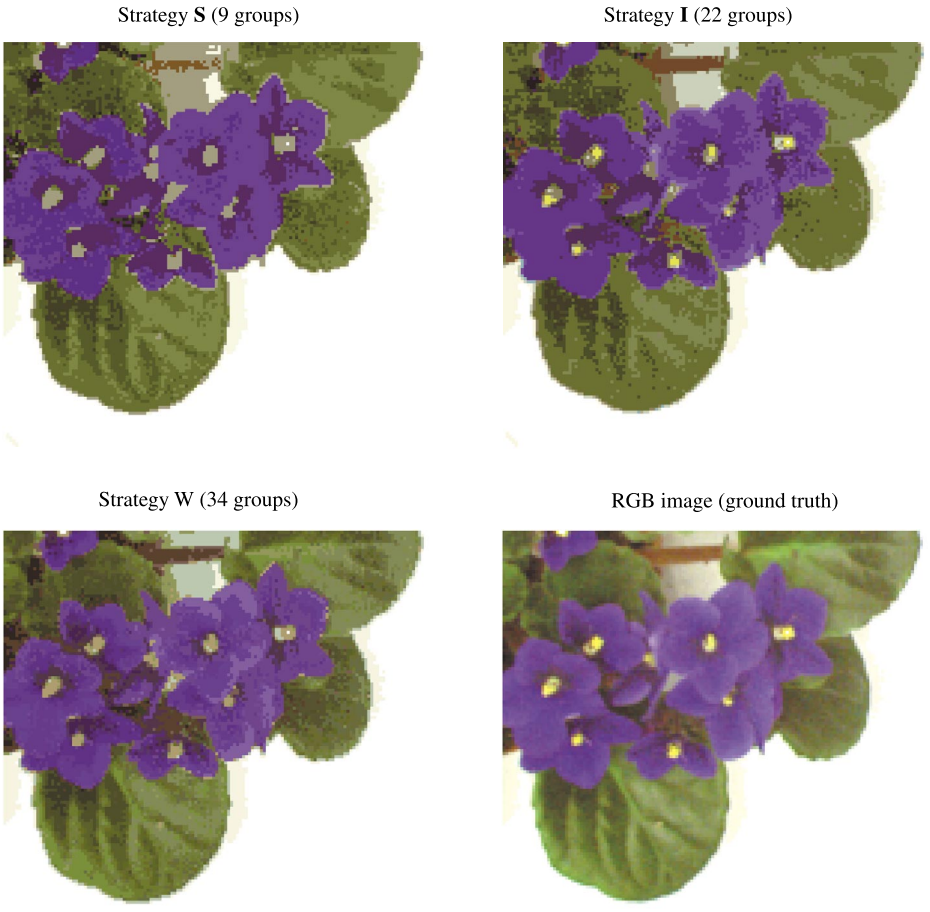


Figure 5. Results for the model-based clustering strategies on the flower image for the same random initial sample. The colors for the clusters were obtained by using the mean RGB value for each group. The yellow flower centers are clearly visible in the segmentation produced by Strategy **I***, but not in those produced by strategies* **S** *and* **W**.

summarized in Table 2. By design, both the BIC values and numbers of clusters are larger for Strategy **I** than for Strategy **S**. In fact, Strategy **I** yielded models with substantially more clusters and higher BIC values than Strategy **S**. Moreover, while none of the segmentations from Strategy **S** revealed the yellow flower centers, they were clearly visible in the segmentations produced by Strategy **I** for 99 out of 100 random initial samples. A representative example is shown in Figure 5.

Details of the iterations for Strategy **I** for the sample corresponding to the results shown in Figure 5 are shown in Table 3. In these iterations, Strategy **S** chooses a nine-component `VEV` model in which the mixture components or clusters share a common shape. The number of clusters increases by one at each iteration except at the ninth iteration when BIC can no longer be increased by adding a component in the `VEV` model and the model changes to the unconstrained model `VVV`. Note that the numbers of observations in the smallest and largest clusters are not monotonic functions of the iteration number.

Table 3. Iterations of Incremental Model-Based Clustering for the St. Paulia Flower Data. 1,188,000 was added to all the BIC values. "Smallest cluster" is the number of pixels in the smallest cluster; similarly for "largest cluster."

Iteration	*Model*	*No. of clusters*	*Shifted BIC*	*Smallest cluster*	*Largest cluster*
0	VEV	9	−9186	726	11674
1	VEV	10	−7531	708	12104
2	VEV	11	−7380	450	11978
3	VEV	12	−7051	389	12128
4	VEV	13	−7024	142	12124
5	VEV	14	−4786	555	12785
6	VEV	15	−4696	339	12772
7	VEV	16	−3797	282	12170
8	VEV	17	−3064	318	12291
9	VVV	17	−1710	305	7615
10	VVV	18	−1627	141	7484
11	VVV	19	−1313	66	7308
12	VVV	20	−1254	65	7301
13	VVV	21	−867	65	7281
14	VVV	22	−860	66	7297

3.3 BRAIN MRI DATA

These data describe a four-band MRI of a brain with a tumor, shown in gray scale in Figure 6. [These data were obtained from Professor A. Heerschap of the Radboud University Medical Center, Nijmegen, The Netherlands.] There are 23,712 pixels.

Results for the model-based clustering strategies on the brain MRI (Figure 6) are summarized in Table 4. As for the flower data, Strategy **I** leads to models with more clusters and higher BIC values than does Strategy **S**. Clusterings for the three model-based clustering strategies for the same random initial sample are shown in Figure 7. In this case the tumor is large enough to be detected by Strategy **S** with sample size 2,000. One thing to note is that, for Strategy **I**, the light blue area at the periphery of the tumor region is confined to the tumor area, while for Strategy **S** it is scattered over other regions of the brain. Another observation is that the segmentation from Strategy **I** is smoother and more appealing to the eye than the fragmented segmentation from Strategy **W**.

4. DISCUSSION

Incremental model-based clustering is a conceptually simple EM-based method that finds small clusters without subdividing the data into a large number of groups. It can be combined with any in-core method for improving performance of EM for large datasets. A further advantage of the incremental approach is that the evolution of the clusters can be monitored and the process stopped or interrupted as clusters of interest emerge. For complex datasets, such as the image data shown here, incremental model-based clustering improves on the ability of the simple sample-based approach to reveal significant small

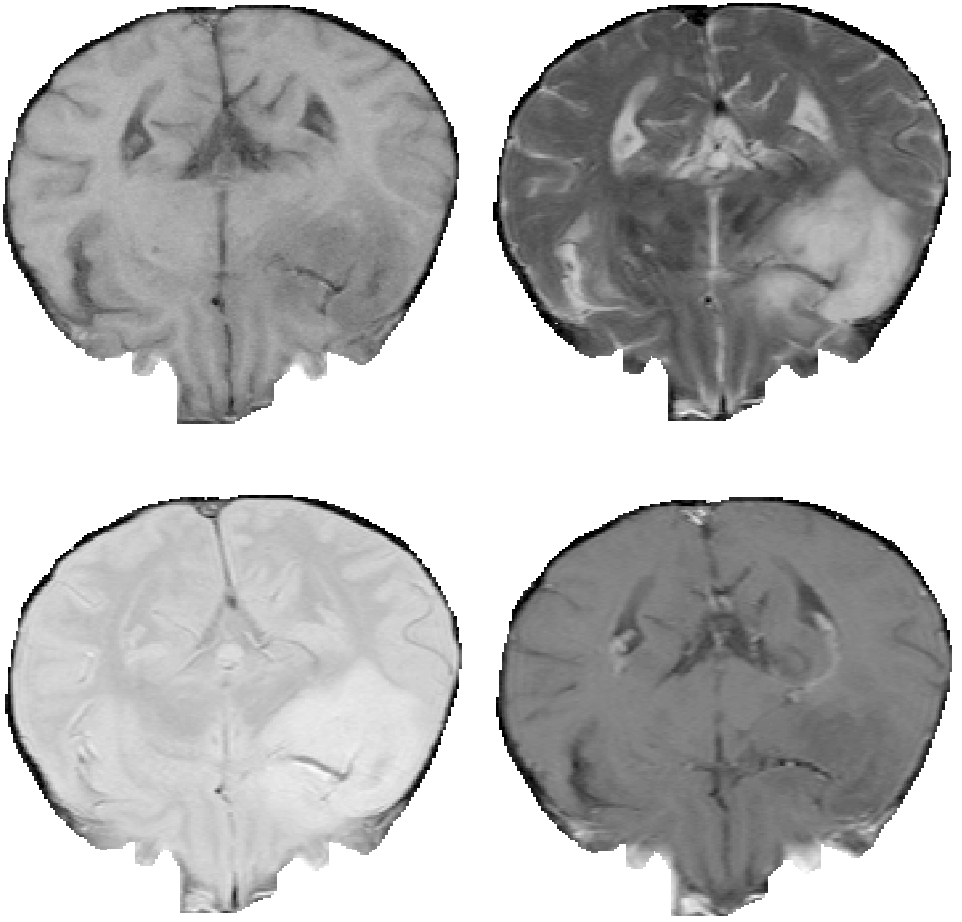


Figure 6. The four bands of the brain MRI shown in gray scale. There are 23,712 pixels. Lesions are evident in the tumor region (at the lower right in each image).

clusters without adding severely to the computational burden, and without overfitting.

It is clear from Tables 2 and 4 that the results do vary with the initial sample. Based on extensive experiments in both real simulated datasets, Wehrens et al. (2004) showed that as long as the initial random sample is not too small, stability of the classification of pixels over different samples is directly reflected in the uncertainty of the probabilistic classification. Recall that EM for mixtures produces a conditional probability that an observation belongs to each component or cluster, rather than a discrete classification.

Table 5 gives a rough indication of the computational effort associated with the different strategies. These timings should not be interpreted as a rigorous algorithmic comparison, because the code has not been optimized for speed, and because it is not possible to separate the timing overhead due to the interpreted R interface (which would be considerable for datasets of this size) from the time it takes to do the actual computations.

Strategy **W** is by far the most expensive strategy in terms of computing time. Strategy

Table 4. Results for the Model-Based Clustering Strategies on the Brain MRI Data. One hundred different random initial samples were used; "sd" denotes the standard deviation. 755,000 was added to all the BIC values.

Strategy	*Mean*	*sd*	*min*	*max*
		Shifted BIC		
S	−6675	1165	−8601	−3794
I	−2418	447	−3825	−1658
W	−867	69	−1025	−670
		Number of Clusters		
S	9.1	1.7	7	14
I	17.6	1.7	13	22
W	33.7	3.5	26	40

I takes somewhat more time than Strategy **S**, but considerably less time than Strategy **W**, while producing better results. It should be kept in mind that timings for Strategy **S** and Strategy **W** depend on the maximum number of components considered.

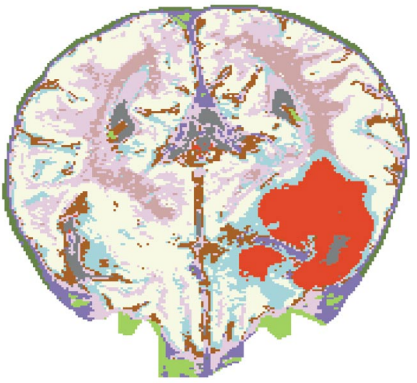
Several other approaches to clustering large datasets based on forming a model from a sample of the data have been proposed. Fayyad and Smyth (1997, 1999) suggested a methodology for discovering small classes in large datasets, in which a model is first constructed based on a sample, and then applied to the entire dataset. Observations that are well-classified by the model are retained along with a stratified sample of the rest of the observations, and the procedure is repeated until all observations are well classified.

Bradley, Fayyad, and Reina (1998) developed a one-pass (exclusive of sampling) method based on EM for mixture models that divides the data in to three classes: records that can be discarded (membership certain), records that can be compressed (known to belong together), and records that must be retained in memory for further processing (membership uncertain). Records in the first two classes are represented by their sufficient statistics in subsequent iterations. Records to be compressed are determined by k-means. Several candidate models can be updated simultaneously. The number of clusters present in the data is assumed to be known in advance.

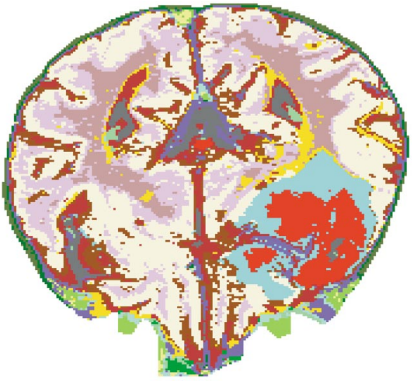
Maitra (2001) proposed a multistage algorithm that clusters an initial sample using a mixture modeling approach, filters out observations that can be reasonably classified by these clusters, and iterates the procedure on the remainder. Sample size can be adjusted to accomodate available computational resources. The method requires only a few stages on very large datasets, but produces many clusters. One reason for the large number of clusters is the assumption that mixture components have a common covariance, a requirement of the hypothesis test used to determine representativeness of the identified clusters.

Of course, any sampling-based strategy can be applied to the data for a number of samples, increasing the chances that a good model for the data will be found. Meek, Thiesson, and Heckerman (2001) gave a strategy for determining sample size in methods that extend a sample-based approach to the full dataset. The basic idea is to apply a training algorithm to larger and larger subsets of the data, until expected costs outweigh the expected benefits associated with training. The premise is that the best results come from the training algorithm applied to all of the data. A decision-theoretic framework for cost-benefit analysis is

Strategy **S** (11 groups)



Strategy **I** (17 groups)



Strategy **W** (32 groups)

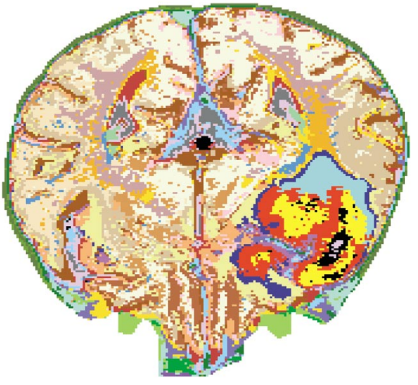


Figure 7. Results for the three model-based clustering strategies on the MRI image for the same random initial sample. For Strategy **I**, *the light blue area around the periphery of the tumor is confined to the tumor region, while for Strategy* **S** *it is dispersed over the image. Strategy* **W** *fits many more components in the tumor region. No spatial information was used in clustering the data.*

Table 5. Approximate Timings (minutes) for the St. Paulia Flower RGB Image and the Brain MRI. The maximum number of components considered is 15 for Strategy **S** and 40 for Strategy **W**. Timing for Strategy **I** includes Strategy **S** for initialization. Fortran code with R interface; Pentium 4 2.4 GHz processor.

	Strategy **S**	*Strategy* **I**	*Strategy* **W**
flower image	1.5	10.4	119.7
brain MRI	1.0	3.4	63.3

proposed, in which cost is given in terms of computation time, and benefit or accuracy is the value of the log-likelihood. Situation and data-dependent scaling issues are discussed. It is assumed that the number of components in the mixture model is known.

There are also several approaches that are not based on a model derived from a sample. DuMouchel et al. (1999) proposed strategies for scaling down massive datasets so that methods for small to moderate sized numbers of observations can be applied. The data are first grouped into regions, using bins induced by categorical variables and bins induced either by quantiles or data spheres for quantitative variables. Moments are then calculated for the elements falling in those regions. Finally, a set of squashed data elements is created for each region, whose moments approximate those of the observations in that region. This produces a smaller dataset to be analyzed which consists of the squashed data elements and induced weights. The squashed data can then be analyzed by conventional methods that accept weighted observations. A potential problem is that small clusters may be missed, although the authors point out that the initial grouping could be constructed in a partially supervised setting to detect small clusters with known characteristics.

Meek et al. (2002) proposed a stagewise procedure that uses reweighted data to fit a new component to the current mixture model. As in our method, the new component is accepted only if it results in an improvement in BIC. Unlike in our method, observations that are not well-predicted in the current model are given more weight in fitting the new component, while previous components remain fixed. The authors point out that the method could be combined with backfitting procedures that could update all components of the model. The method requires an initial estimate for mixing proportions and model parameters of new component at each stage, whereas our method starts each stage with a new classification of the data that separates out observations of low density.

Posse (2001) extended model-based hierarchical clustering to large datasets by using an initial partition based on the minimum spanning tree. Although the minimum spanning tree can be computed quickly, it tends to produce groups that are too large to be useful, so the method is supplemented with strategies for subdividing these groups prior to clustering. Once an observation is grouped in the initial partition, it cannot be separated from that group in hierarchical clustering. Also, there are no methods for choosing the number of clusters in model-based hierarchical clustering that are comparable to those available for mixture models. However, this approach could be combined with the mixture modeling approach as a starting scheme for a strategy similar to Strategy **W**.

Tantrum, Murua, and Stuetzle (2002) extended model-based hierarchical clustering to large datasets through "refractionation," which splits the data up into many subsets or

"fractions." The fractions are clustered by model-based hierarchical clustering into a fixed number of groups and then summarized by their means into meta-observations. These meta-observations are in turn grouped by model-based clustering, after which a single EM step is applied to the conditional probabilities defined by the classifications to approximate a mixture likelihood for computing the BIC, which is used to determine the number of clusters. Initially the data are divided randomly, but in subsequent iterations, clusters larger than a fixed fraction size are split into fractions. Observations are assigned to the cluster with the closest mean, and the procedure is iterated until successive partitions cease to become more similar.

Recently, a number of techniques have been developed for speeding up the EM algorithm on large datasets. Incremental model-based clustering, which relies on the EM algorithm, can be implemented in combination with any of these. Several of these methods are based on a partial E-step. In incremental EM (Neal and Hinton 1998; Thiesson, Meek, and Heckerman 2001), the E-step is updated in blocks of observations. For normal mixtures, the M-step can be efficiently implemented to update sufficient statistics in blocks. Lazy EM (McLachlan and Peel 2000; Thiesson et al. 2001) identifies observations for which the maximum posterior probability is close to 1 and updates the E-step only for the complement of this subset for several iterations. A full E-step must be performed periodically to ensure convergence. In sparse EM (Neal and Hinton 1998), only posterior probabilities above a certain threshold are updated for each observation. In the M-step, only the contribution of the corresponding sufficient statistics need be updated. As for lazy EM, a full E-step needs to be performed periodically to ensure convergence. Moore (1999) organized the data in a multiresolution kd-tree, that allows fast approximations to the E-step in the EM algorithm by eliminating computations that are considered ignorable. The gain in efficiency diminishes as the dimension of the data increases. Another approach for large datasets is a componentwise EM for mixtures (Celeux, Chrétien, Forbes, and Mkhadri 2001) in which parameter estimates are decoupled so as to reduce the size of the missing data space computed in the E-step.

The EM algorithm is well known to have a linear of rate of convergence, which can sometimes be very slow. Redner and Walker (1984) suggested using a few steps of EM to start a maximization procedure with faster asymptotic convergence. Superlinearly convergent methods would certainly be useful in this context from the point of view of reducing the overall number of iterations (each of which is costly due to the amount of data involved), as well as for assurance that a local maximum has indeed been reached (the latter being difficult to determine for slow linearly convergent methods). Although robust implementations of such methods are not available in the public domain, good software is available commercially (see, e.g., http://www-fp.mcs.anl.gov/otc/Guide/SoftwareGuide/index.html). State-of-the-art methods can easily handle problems with large numbers of parameters such as those that arise here, provided good starting values are available. The objective and constaints are specified using a modeling language such as AMPL (Fourer, Gay, and Kernighan 2002), which allows automatic differentiation that exploits sparsity and reuse of computational expressions.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grant 5 R01 EB002137–03 and by Office of Naval Research contract N00014–01–10745. The authors thank Christophe Ambroise for thoughtful discussion of this work at the 2003 Session of the International Statistical Institute in Berlin.

[Received January 2004. Revised June 2004.]

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Bradley, P. S., Fayyad, U., and Reina, C. (1998), "Scaling Clustering Algorithms to Large Databases," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, Menlo Park, CA: AAAI Press, pp. 9–15.

Campbell, J. G., Fraley, C., Murtagh, F., and Raftery, A. E. (1997), "Linear Flaw Detection in Woven Textiles Using Model-Based Clustering," *Pattern Recognition Letters*, 18, 1539–1548.

Campbell, J. G., Fraley, C., Stanford, D., Murtagh, F., and Raftery, A. E. (1999), "Model-Based Methods for Real-Time Textile Fault Detection," *International Journal of Imaging Systems and Technology*, 10, 339–346.

Celeux, G., Chrétien, S., Forbes, F., and Mkhadri, A. (2001), "A Component-Wise EM Algorithm for Mixtures," *Journal of Computational and Graphical Statistics*, 10, 699–712.

Celeux, G., and Govaert, G. (1995), "Gaussian Parsimonious Clustering Models," *Pattern Recognition*, 28, 781–793.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood for Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999), "Squashing Flat Files Flatter," in *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, New York: ACM Press, pp. 6–15.

Fayyad, U., and Smyth, P. (1997), "From Massive Datasets to Science Catalogs: Applications and Challenges," in *Statistics and Massive Data Sets: Report to the Committee on Applied and Theoretical Statistics*, eds. J. Kettenring and D. Pregibon, National Research Council.

——— (1999), "Cataloging and Mining Massive Datasets for Science Data Analysis," *Journal of Computational and Graphical Statistics*, 8, 589–610.

Fourer, R., Gay, D. M., and Kernighan, B. W. (2002), *AMPL: A Modeling Language for Mathematical Programming* (2nd ed.), Belmont, CA : Duxbury Press/Brooks/Cole Publishing Company.

Fraley, C. (1998), "Algorithms for Model-Based Gaussian Hierarchical Clustering," *SIAM Journal on Scientific Computing*, 20, 270–281.

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.

——— (1999), "MCLUST: Software for Model-Based Cluster Analysis," *Journal of Classification*, 16, 297–306.

——— (2002), "Model-Based Clustering, Discriminant Analysis and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

——— (2003), "Enhanced Software for Model-Based Clustering, Density Estimation, and Discriminant Analysis: MCLUST," *Journal of Classification*, 20, 263–286.

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, eds. L. M. L. Cam and J. Neyman, Volume 1, University of California Press, pp. 281–297.

Maitra, R. (2001). "Clustering Massive Datasets with Applications in Software Metrics and Tomography," *Technometrics*, 43, 336–346.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models : Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

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

Meek, C., Thiesson, B., and Heckerman, D. (2001), "The Learning-Curve Sampling Method Applied to Model-Based Clustering," *Journal of Machine Learning Research*, 2, 397–418. Also in *AI and Statistics* 2001.

——— (2002), "Staged Mixture Modeling and Boosting," in *Proceedings of Eighteenth Conference on Uncertainty in Artificial Intelligence,* Edmonton, Alberta, San Francisco: Morgan Kaufmann, pp. 335–343.

Moore, A. (1999), "Very Fast EM-Based Mixture Model Clustering Using Multiresolution *kd*-Trees," in *Advances in Neural Information Processing Systems* (vol. 11), eds. M. Kearns, S. Solla, and D. Cohn, Cambridge, MA: MIT Press, pp. 543–549.

Mukherjee, S., Feigelson, E. D., Babu, G. J., Murtagh, F., Fraley, C., and Raftery, A. E. (1998), "Three Types of Gamma Ray Bursts," *The Astrophysical Journal*, 508, 314–327.

Neal, R., and Hinton, G. (1998), "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants," in *Learning in Graphical Models*, ed. M. Jordan, San Francisco: Klüwer, pp. 355–371.

Posse, C. (2001), "Hierarchical Model-Based Clustering for Large Datasets," *Journal of Computational and Graphical Statistics*, 10, 464–486.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Tantrum, J., Murua, A., and Stuetzle, W. (2002), "Hierarchical Model-Based Clustering of Large Datasets Through Fractionation and Refractionation," in *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press, pp. 183–190.

Thiesson, B., Meek, C., and Heckerman, D. (2001), "Accelerating EM for Large Datasets," *Machine Learning*, 45, 279–299.

Wang, N., and Raftery, A. E. (2002), "Nearest Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest Neighbor Cleaning" (with discussion), *Journal of the American Statistical Association*, 97, 994–1019.

Ward, J. H. (1963), "Hierarchical Groupings to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 234–244.

Wehrens, R., Buydens, L., Fraley, C., and Raftery, A. (2004), "Model-Based Clustering for Image Segmentation and Large Datasets via Sampling," *Journal of Classification*, 21, 231–253.

Wehrens, R., Simonetti, A., and Buydens, L. (2002), "Mixture-Modeling of Medical Magnetic Resonance Data," *Journal of Chemometrics*, 16, 1–10.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001), "Model-Based Clustering and Data Transformation for Gene Expression Data," *Bioinformatics*, 17, 977–987.