Probabilistic Clustering using Hierarchical Models

Technical Report No. 99–16
Department of Information and Computer Science
University of California, Irvine

Igor Cadez and Padhraic Smyth
Department of Information and Computer Science
University of California, Irvine
CA 92697-3425
[icadez,smyth@ics.uci.edu]

March 1999

Abstract

This paper addresses the problem of clustering data when the available data measurements are not multivariate vectors of fixed dimensionality. For example, one might have data from a set of medical patients, where for each patient there are time series, image, text, and multivariate data. We propose a general probabilistic clustering framework for clustering heterogeneous data types of this form. We focus on two-level probabilistic hierarchical models, consisting of a high-level mixture model on parameters and a low-level model for observations. This general framework permits probabilistic clustering of "objects" (sequences, histograms, images, etc) using an extension of the expectation-maximization (EM) algorithm which we derive. We further show that earlier (intuitive) clustering algorithms can be viewed as special cases (approximations) of the framework proposed here. The paper includes several illustrations of the method, including an application to a problem in clustering two-dimensional histograms of red blood cell data in a medical diagnosis context.

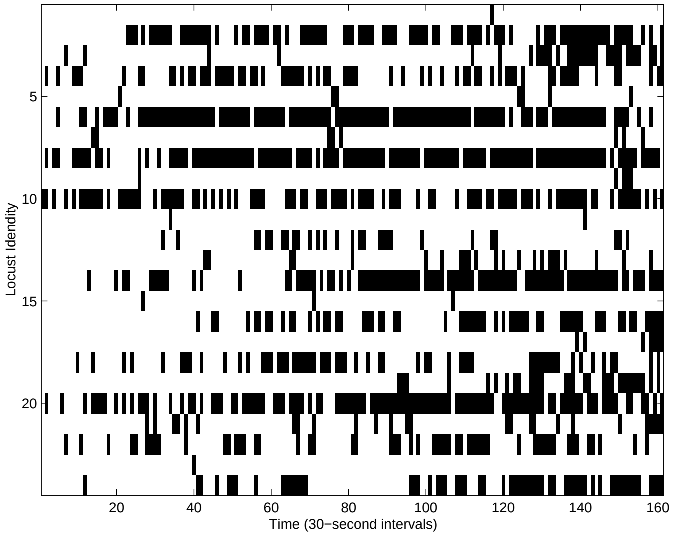

Figure 1: Binary activity data from 24 locusts as a function of time (white indicates inactive, black indicates active. The odd numbered (fed) locusts are less active (white) than the even numbered (unfed) locusts.

1 Introduction

Clustering is a fundamental and widely applied methodology in understanding large data sets. Clustering algorithms are typically applied to vector measurements of fixed dimension. For example, we may have a d measurements on a set of patients and we represent the measurements on individual i as a d-dimensional vector. For such data there are a wealth of different clustering methods available.

Increasingly, however, data may not be in a convenient fixed-dimensional form. For example, in a medical context, one might have data such as biomedical time series, genetic sequences, diagnostic images, text annotations describing a diagnosis, as well as "standard" vector measurements describing age, weight, test results, and so forth. Complicating matters is the fact that one could have different lengths of sequences and time series or different sized images for different individuals, as well as different numbers of sequences or images per individual.

Such *heterogeneous* data sets are increasingly common, not only in medicine, but in many other fields. For example, in transaction data (such as Web logs, retail purchases, etc) one can have temporal, spatial, and multivariate aspects to an individual's measurements. Scientific studies also generate multiple data types. For example, in cognitive science and animal behavioral studies it is common to have behavioral (dynamic, sequential) data for each individual. Figure 1 shows a simple example of such data, where binary observation sequences were collected on 24 locusts. Here one would like to be able to cluster the locusts based on the observed behavioral sequences.

The fundamental problem in clustering such data lies in defining a suitable *distance* between any pair of individuals. For multivariate data we can work from geometric notions of distance in a Euclidean space (e.g., the k-means algorithm). For data composed of sequences, images, and so forth, there need not be any obviously equivalent geometric analogies.

Traditionally, two general techniques have been used to address this problem. The first approach reduces all of the heterogeneous data to a fixed-length vector representation. For example, the vector-space representation is widely used in information retrieval for representing text documents of different lengths and formats, and histograms are often used to summarize Web access sequences or the color content of pixel images. In practice, this technique is often very useful and has the significant advantage of being relatively simple. It allows one to take full advantage of the many vector-space clustering algorithms. However, it is also clear that this general approach incurs some information loss (e.g., the loss of dynamic information when summarizing sequential data with marginal statistics), which in some applications may be critical. From a knowledge discovery perspective the vector representation may be adequate for clustering but can nonetheless be a relatively weak representation of the semantic content of the data.

A second general approach to clustering object data is to directly define a pairwise distance function between objects, typically based on some notion of how easy it is to "transform" one object into the other. A well-known example here is the *edit distance* between sequences. For example, in computational biology relatively sophisticated edit distances for pairs of protein sequences have been developed based on first principles knowledge of how such sequences evolve genetically. Given N objects, we generate a matrix of N^2 pairwise distances (so-called proximity data). Once again there are a large number of algorithms for clustering based on proximity data (e.g., hierarchical clustering). However, for large data sets the $O(N^2)$ cost of constructing the proximity matrix is prohibitive, although it always possible to work with a sampled version of this matrix. In addition, proximity data clustering is not always ideal from a practical viewpoint. For example, there is no natural probabilistic setting for proximity data clustering, and thus, no natural setting for answering questions such as "how many clusters?" in a statistical context.

In this paper we investigate the general question of how one can cluster individuals based on heterogeneous data and investigate a general probabilistic hierarchical clustering framework for this problem.

2 Probabilistic Model-Based Clustering

2.1 Model-Based Clustering of Multivariate Data

Before considering modeling of heterogeneous data, it is useful to first review the basic concepts of probabilistic clustering for "standard" vector (multivariate) data. A *generative probabilistic model* is a probabilistic model for how an observed data point $\mathbf{x}$ is generated. For example, the model could be a multivariate Gaussian $f(\mathbf{x}|\theta)$, where θ are the parameters of the Gaussian model. It is important to keep in mind that this technique is model-based, i.e., we are hypothesizing a data-generating mechanism (in the form of the density function f) for the observed data.

Probabilistic model-based clustering (e.g., see McLachlan and Basford (1988), Stutz and Cheeseman (1996), Smyth (1996), and Fraley and Raftery (1998)) builds on this idea by hypothesizing that the data are being generated by K different models. These K models correspond to the K clusters of interest. A particularly useful model of this form is the *finite mixture model*, which has the general form:

$$f(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}|\theta_k)\alpha_k \qquad (1)$$

where the f_k's are the density functions (locations and shapes in d-dimensional space) for each cluster, and the α_k are the relative weights (or probability) of each cluster, $\sum \alpha_k = 1$. The parameters can be estimated from the observed data using the expectation maximization (EM) algorithm which we discuss in detail later.

This form of clustering can have distinct advantages over competing non-probabilistic approaches (such as the K-means algorithm) for certain problems, since it allows uncertainty in cluster membership, direct control over the variability allowed within each cluster (as captured by the variance characteristics of each component model f_k), and permits an objective treatment of the ever-thorny question of how many clusters are being suggested by the data (for example, see Smyth et al (1997) and Smyth, Ide, and Ghil (in press) for a specific application in atmospheric science).

2.2 A Generative Hierarchical Cluster Model

We propose the following hierarchical framework for probabilistic clustering:

- An individual is randomly drawn from the overall population (universe) and is indexed with letter i (the index represents the "individual count").
- The individual is assigned to one of K clusters with probability α_k, $\sum \alpha_k = 1$, $1 \leq k \leq K$.
- Parameters Θ_i are drawn from a "prior" model in parameter space given membership in cluster k, namely $\pi_k(\Theta_i|\phi_k)$, where ϕ_k are the prior parameters for Θ in the kth cluster. This prior represents the *within-cluster variability* in parameter space of individuals belonging to cluster k. *Between-cluster* differences are obtained from having different prior components with different prior parameters ϕ_k for each of the K clusters. We will refer to this model on parameters Θ_i as the *high level model.* If we do not know which cluster an individual's Θ_i's come from then the high level model is a mixture
$$\pi(\Theta) = \sum_{k=1}^{K} \alpha_k \pi_k(\Theta|\phi_k).$$
- Given Θ_i, data D_i are now generated by a data generating probability model $f(D_i|\Theta_i)$. We refer to this as the *low level model.* D_i represents the heterogeneous data associated with individual i. Note that this data can be quite general in form, e.g., sequences, images, histograms, text, vectors, or combinations of any of these, as long as we can define a density function f on such data.

This type of hierarchical model is quite plausible as a generative mechanism for a wide variety of clustering situations. For example, an individual (call them i) might be predisposed to contract a disease for hereditary reasons with probability α_k. The disease might manifest itself via unusual EEG patterns. The EEG time series could be governed by some parameters Θ_i, i.e., there is a density function $f(D_i|\Theta_i)$ on observed time series values D_i for this individual. In turn, there is a distribution of Θ's for the population of individuals

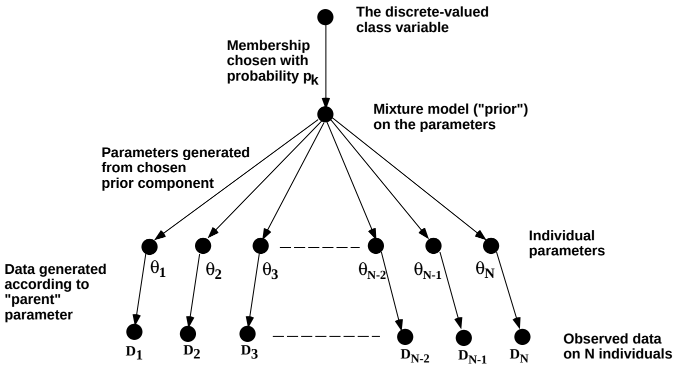
The discrete-valued class variable
Membership chosen with probability p_k
Mixture model ("prior") on the parameters
Parameters generated from chosen prior component
Individual parameters
Data generated according to "parent" parameter
θ_1
θ_2
θ_3
θ_{N-2}
θ_{N-1}
θ_N
D_1
D_2
D_3
D_{N-2}
D_{N-1}
D_N
Observed data on N individuals

Figure 2: An illustration of the hierarchical cluster model.

as a whole for this disease, and the typicality and variation in the Θ's (and indirectly in the observed time series) is reflected in the prior $\pi_k(\Theta|\phi_k)$ for this component.

The hierarchical model allows us to model heterogeneous data across different individuals in a fairly general framework. Using the model, our goal will be to estimate the ϕ_k (cluster characteristics) and Θ_i (individual parameters), given only observed data D_i. Figure 2 illustrates the structure of the overall hierarchical model, and can be interpreted as a directed graphical model (belief network). Low-level individual parameters and data are assumed conditionally independent given the high-level model.

As we will see in this paper, the probabilistic hierarchical framework has some distinct advantages over alternative methods for object clustering. For example, clustering sequences of different lengths is not problematic. The key idea is that objects are defined to be similar in terms of common similarity to a model, expressed through the likelihood function $f(D_i|\Theta)$. Two objects are considered similar if they both have higher likelihood under one particular model than under any other model. For example, a short sequence *abaabab* and a longer sequence *baababbabbabaabaab* could both be considered similar in the context of two alternative Markov models, one of which favors very short runs of a or b (the more likely model for these two sequences), and one of which favors much longer runs.

This general type of model is commonly used in statistics, and is variously referred to as a hierarchical model or a multilevel model (e.g., Gelman et al, 1994). Thus, the model itself is not new. What is new in this paper is the use of this type of model for clustering, where we specifically make the high-level model have a finite mixture form. Note this is not a conventional prior in the sense that the parameters of this prior Φ are learned from the data (this is quite similar to a technique known as "empirical Bayes"). We are not aware of any related published work in probabilistic model-based clustering which uses this general framework, i.e., a hierarchical model with a mixture model prior which is specifically being used for clustering. For example, Sjolander et al (1996) used mixtures of Dirichlet priors for modeling protein sequences but with a focus on prediction rather than clustering. Their framework can be viewed as a special case of the framework we set out below.

2.3 Notation

The observed data is $D = \{D_1, D_2, \ldots, D_n\}$, where each D_i is a *set* of observations for individual i. For example, each D_i could be a set of time series or sequence values, a set of pixels, a set of histogram counts, and so forth. The only requirement is that we can define a probabilistic model (the low-level model) for the observed data as $f(D_i|\Theta_i)$. The set of all low level model parameters is $\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_n\}$. The low level parameters Θ are modeled by the high level mixture model $\pi(\Theta) = \sum_{k=1}^{K} \alpha_k \pi_k(\Theta|\phi_k)$, where the model is parameterized by $\Phi = \{\alpha_k, \phi_k\}, 1 \leq k \leq K$. Note that the mixture is a consequence of missing (hidden) data about which group an individual belongs to.

We use D to represent all of the observed data, Z to represent all of the missing data, Θ to represent the low level model parameters, and Φ to represent the high level model parameters. The missing data Z typically consists of missing class labels k_i (from the generative model), but can include any additional missing data that might appear in the specific low level model $f(D|\Theta)$ (e.g., when this model is also a mixture). Accordingly, we split the missing data Z into two subsets $Z = \{Z_k, Z_r\}$ where Z_k represents the missing data in the high level model, while Z_r represents the missing data in the low level model.

3 EM for Standard and Hierarchical Models

3.1 Standard EM

The EM algorithm is a convenient way of maximizing the log-likelihood associated with mixture models. Assume that there is some observed data X, some missing data Z and that the joint distribution of X and Z is $p(X, Z|\theta)$, where θ is the set of all the parameters. The marginal distribution of X is $p(X|\theta) = \int p(X, Z|\theta) dZ$. Define the log-likelihood associated with each of these two distributions as

$$\begin{aligned} l_{X,Z}(\theta) &= \log p(X, Z|\theta) \\ l_X(\theta) &= \log p(X|\theta) = \log \int p(X, Z|\theta) dZ \end{aligned}$$

The goal is to maximize the log-likelihood of the observed data, $l_X(\theta)$. If we decompose the joint as

$$p(X, Z|\theta) = p(X|\theta) p(Z|X, \theta),$$

take the log of the both sides of the equation and rearrange the terms, we get:

$$l_X(\theta) = l_{X,Z}(\theta) - \log p(Z|X, \theta).$$

This equation can be averaged over any choice of PDF $f(Z)$, but if we make the particular choice of $f(Z) = p(Z|X, \theta')$, we get:

$$\begin{aligned} l_X(\theta) &= \langle l_{X,Z}(\theta) \rangle_{p(Z|X,\theta')} - \int p(Z|X, \theta') \log p(Z|X, \theta) \\ &= Q(\theta, \theta') - h(\theta, \theta') \end{aligned}$$

The left hand side of the equation is unchanged by the averaging since it does not depend on Z (it also represents the likelihood that we want to maximize). The first term on the right hand side represents the *expected value of the full likelihood* $Q(\theta, \theta')$ (calculating Q represents the so called "E" step), while the second term represents the negative cross-entropy $h(\theta, \theta')$ and has the desirable property that it is maximized by $\theta = \theta'$. The property of the cross entropy term gives rise to the EM algorithm where one maximizes $Q(\theta, \theta')$ with respect to θ (the so called "M" step) by using the current value of parameters for θ'. By iteratively applying E and M steps the likelihood on the *lhs* of the equation is guaranteed not to decrease and will reach the local maximum under some very weak assumptions. Note that $Q(\theta, \theta')$ need not be maximized at each iteration; it suffices to find θ' that will increase $Q(\theta, \theta')$. Algorithms that exploit this fact are known as generalized EM algorithms (GEM) (e.g., McLachlan and Krishnan, 1997).

3.2 EM for Hierarchical Cluster Models

In the hierarchical model, we have data D, hidden data Z (consisting of Z_k — class labels k_i for each individual i, and Z_r — additional low level hidden data), a set of parameters Θ, and a set of high level parameters Φ.

We treat both Φ and Θ_i as parameters and seek the maximum a posteriori (MAP) estimate in the following manner:

$$p(D|\Theta, \Phi)p(\Theta, \Phi) = p(D, \Theta|\Phi)p(\Phi).$$

We will assume that the hyperprior $p(\Phi)$ is uninformative so we omit it from further analysis and write the optimization problem as:

$$\{\hat{\Theta}, \hat{\Phi}\} = \arg\max_{\Theta,\Phi} p(D, \Theta|\Phi). \tag{2}$$

Alternatively one could follow a fully Bayesian approach by including the hyperprior and by integrating over unknown parameters rather than obtaining point estimates; in this paper we will restrict our attention to a MAP analysis.

3.3 Optimization for hierarchical models

As the first step we need to show that the EM algorithm can be applied to solving equation (2). Note that one of the variables we are optimizing with respect to (i.e. Θ) appears as a variable in the joint and not as a conditioning variable. If we decompose the joint as:

$$p(D, \Theta, Z|\Phi) = p(Z|D, \Theta, \Phi)p(D, \Theta|\Phi),$$

the derivation of EM is similar to the standard derivation:

$$\begin{aligned} l_{D,\Theta}(\Phi) &= l_{D,\Theta,Z}(\Phi) - \log p(Z|D, \Theta, \Phi) \\ l_{D,\Theta}(\Phi) &= \langle l_{D,\Theta,Z}(\Phi)\rangle_{p(Z|D,\Theta',\Phi')} - \int p(Z|D, \Theta', \Phi') \log p(Z|D, \Theta, \Phi) dZ. \\ &= Q(\Theta, \Phi; \Theta', \Phi') - h(\Theta, \Phi; \Theta', \Phi') \qquad (3) \end{aligned}$$

From Equation (3) it is obvious that the last term still represents the cross-entropy, hence the EM algorithm is still valid. The equation also establishes the specific form of the Q function.

We now calculate all the relevant probabilities that appear in the $Q(\Theta, \Phi; \Theta', \Phi')$. The joint distribution of observed and hidden data can be inferred directly from the generative model:

$$\begin{aligned} p(D_i, \Theta_i, Z_i|\phi_{k_i}) &= f(D_i, Z^i_D|\Theta_i)\alpha_{k_i}\pi_{k_i}(\Theta_i|\phi_{k_i}) \\ p(D, \Theta, Z|\Phi) &= \prod_{i=1}^{n} f(D_i, Z^i_r|\Theta_i)\alpha_{k_i}\pi_{k_i}(\Theta_i|\phi_{k_i}), \qquad (4) \end{aligned}$$

under an iid assumption (not necessary in general). Here we explicitly include the dependence on any missing data Z_r in the low level model. The missing data Z_r naturally divides into subsets for each individual: $Z_r = \{Z_r^1, Z_r^2, \ldots, Z_r^n\}$. The corresponding log-likelihood is:

$$l_{D,\Theta,Z}(\Phi) = \sum_{i=1}^{n} \log\left[\alpha_{k_i} \pi_{k_i}(\Theta_i|\phi_{k_i}) f(D_i, r_i|\Theta_i)\right]. \tag{5}$$

Note that class labels k_i for each individual represent the missing data Z_k in the high level model. We also explicitly write ϕ_{k_i} to show that each mixture component has its own set of parameters (i.e., the mixture parameters for different components are independent). The conditional probability of missing data given the observed data is

$$\begin{aligned} p(Z|D,\Theta,\Phi) &= p(D,\Theta,Z,\Phi)/p(D,\Theta,\Phi) && (6)\\ &= \prod_{i=1}^{n} \frac{\alpha_{k_i}\pi_{k_i}(\Theta_i|\phi_{k_i}) f(D_i, r_i|\Theta_i)}{\pi(\Theta_i|\Phi) f(D_i|\Theta_i)}, && (7) \end{aligned}$$

where we used the index r_i to enumerate the missing data Z_r for the ith individual in the low level model. Note that in the special case when there is no hidden data at the low level, we have that $f(D_i, r_i|\Theta_i) = f(D_i|\Theta_i)$, and the low level model cancels out of the equation to yield:

$$p(Z|D,\Theta,\Phi) = \prod_{i=1}^{n} \frac{\alpha_{k_i}\pi_{k_i}(\Theta_i|\phi_{k_i})}{\pi(\Theta_i|\Phi)}.$$

This form is particularly useful if the low level model is a simple model (e.g., a single Gaussian or a single Markov chain). In general one can derive that the Q function is

$$\begin{aligned} Q(\Theta,\Phi;\Theta',\Phi') &= \langle l_{D,\Theta,Z}(\Phi)\rangle_{p(Z|D,\Theta',\Phi')}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r_i} \frac{\alpha'_k \pi_k(\Theta'_i|\phi'_k) f(D_i, r_i|\Theta'_i)}{\pi(\Theta'_i|\Phi') f(D_i|\Theta'_i)} \log\left[\alpha_k \pi_k(\Theta_i|\phi_k) f(D_i, r_i|\Theta_i)\right]\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r_i} w_{kr}^i \log\left[\alpha_k \pi_k(\Theta_i|\phi_k) f(D_i, r|\Theta_i)\right], && (8) \end{aligned}$$

and the weights w_{kr}^i are naturally defined as:

$$w_{kr}^i = \frac{\alpha'_k \pi_k(\Theta'_i|\phi'_k) f(D_i, r_i|\Theta'_i)}{\pi(\Theta'_i|\Phi') f(D_i|\Theta'_i)}. \tag{9}$$

This equation is the natural generalization of the usual "membership" probabilities in non-hierarchical mixture models, with weights now defined over high-level missing data (the k's) as well as low-level missing data (the r's).

For further simplification of notation, the weights as defined by the previous equation can be interpreted as,

$$\begin{aligned} W_k^i &\equiv \sum_{r_i} w_{kr}^i = \frac{\alpha'_k \pi_k(\Theta'_i|\phi'_k)}{\pi(\Theta'_i|\Phi')}, && (10)\\ w_r^i &\equiv \sum_{k=1}^{K} w_{kr}^i = \frac{f(D_i, r|\Theta'_i)}{f(D_i|\Theta'_i)}, && (11) \end{aligned}$$

where lowercase w and uppercase W represent weights for low and high level models respectively had we treated each level independently. In addition, the summation over both of the lower indices of w^i_{kr} yields:

$$\sum_{k=1}^{K} \sum_{r_i} w^i_{kr} = 1 \tag{12}$$

for any individual i.

Equation (8) represents the "E" step of the EM algorithm (i.e., it is the expected value of the full, unobserved likelihood with respect to the conditional distribution of missing data). All the weights as defined by Equations (9), (10) and (11) depend on the "primed" parameters and are not subject to maximization in the M step.

Assume for the moment that the Θ's are fixed at some particular values and we wish to estimate Φ. We first maximize Q with respect to α (note that there is a constraint $\sum_k \alpha_k = 1$):

$$\frac{\partial}{\partial \alpha_l}\left(Q(\Theta, \Phi; \Theta', \Phi') - \lambda \sum_{k=1}^{K} \alpha_k\right) = \frac{1}{\alpha_l} \sum_{i=1}^{n} \sum_{r_i} w^i_{lr} - \lambda = 0,$$

where upon summing over all l and using property (12) and the definition (10) we readily get:

$$\alpha_l = \frac{1}{n} \sum_{i=1}^{n} \sum_{r_i} w^i_{lr} = \frac{1}{n} \sum_{i=1}^{n} W^i_l, \qquad , 1 \leq l \leq K \tag{13}$$

This is an intuitive result: the weight of each mixture component is determined by the average weight over individuals. Next we maximize Q with respect to the rest of the high level parameters Φ:

$$\begin{aligned} \frac{\partial}{\partial \phi_l} Q(\Theta, \Phi; \Theta', \Phi') &= \sum_{i=1}^{n} \left(\sum_{r_i} w^i_{lr}\right) \frac{\partial}{\partial \phi_l} \log \pi_l(\Theta_i | \phi_l) \\ &= \sum_{i=1}^{n} W^i_l \frac{\partial}{\partial \phi_l} \log \pi_l(\Theta_i | \phi_l) = 0, \end{aligned} \tag{14}$$

which is just a weighted ML estimate of the parameter ϕ_l treating the (fixed) Θ_i as data. Note that this is what one would calculate in the standard EM algorithm if there were no low level model at all.

The previous equations show that the M step at the *high level* is unchanged by the presence of the low level model, if the low level parameters are treated as fixed. However, the low level model will generally change the parameters Θ_i that the high level model treats as input data. Hence, at each iteration, the parameters Θ_i must first be updated as required by the low level model, and only then can the upper procedure be applied to update Φ. To update the parameters Θ_i we write the corresponding M-step equations, but the exact details will depend on the specific low level model being used. Again, we need to find the derivative of Q with respect to each of the parameters Θ_i:

$$\frac{\partial}{\partial \Theta_i} Q(\Theta, \Phi; \Theta', \Phi') = \frac{\partial}{\partial \Theta_i} \sum_{k=1}^{K} \sum_{r_i} w^i_{kr} \log \pi_k(\Theta_i | \phi_k) f(D_i, r_i | \Theta_i)$$

$$= \frac{\partial}{\partial \Theta_i} \sum_{k=1}^{K} \left(\sum_{r_i} w^i_{kr}\right) \log \pi_k(\Theta_i | \phi_k) + \frac{\partial}{\partial \Theta_i} \sum_{r_i} \left(\sum_{k=1}^{K} w^i_{kr}\right) \log f(D_i, r_i | \Theta_i)$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \Theta_i} \sum_{k=1}^{K} W_k^i \log \pi_k(\Theta_i|\phi_k) + \frac{\partial}{\partial \Theta_i} \sum_{r_i} w_r^i \log f(D_i, r_i|\Theta_i) \\
&= 0. \qquad (15)
\end{aligned}
$$

The solution to this equation represents the standard MAP estimate of Θ_i. The prior is defined by the first term in the equation. It consists of a weighted sum of high level mixture components. The second term in the equation is the standard weighted ML term that would appear in an ML estimation at the low level if there were no hierarchical model (i.e., no prior at the high level).

3.4 The 2-level GEM for Hierarchical Models

The results in the previous sections demonstrated that there is a well defined MAP estimation problem for hierarchical models which can be solved by an EM algorithm. The algorithm is a *generalized* EM (GEM) algorithm as the Q function is not maximized at each step. Instead, it is partially maximized twice. Here we summarize the steps:

- Define the hierarchical model by specifying the low level model and the high level model.
- Initialize the algorithm:
  1. Find ML estimates $\hat{\Theta}_{iML}$ of parameters Θ_i for each individual;
  2. Use parameters $\hat{\Theta}_{iML}$ as the data (input) for the high level model and find ML estimates $\hat{\Phi}_{ML}$ of the hyperparameters Φ.

  Each of the two initialization steps can be done independently by using (for example) the standard (decoupled) EM algorithm.
- Iterate through the following steps until convergence, or until the maximum number of iterations has been reached.
  1. Calculate weights w_{kr}^i for each individual using the current parameters and equation (9).
  2. Use the MAP equation (15) to update the parameters Θ_i for each individual.
  3. Use the ML equations (13) and (14) to update the hyper-weights and the rest of the hyperparameters.

The computational complexity will typically be dominated by step 2 above, which takes $O(|D|R)$ steps per iteration, where $|D|$ is the total number of data points observed across all individuals, and R is the number of hidden variables (e.g., mixture components) in the low-level model. The number of GEM iterations before convergence will depend on each data set, but is not unusual for the algorithm to converge within 20 to 50 iterations.

4 Experimental Results and Special Cases of the General Method

In this section we derive some approximations to the full framework and show how the hierarchical models can be used in practice. The most general approach as described by equations (8), (13), (14) and (15) is not necessarily the most useful approach in practice. We look at two specific cases: (a) when there is relatively little variation in parameters within a cluster, and (b) when there is a relatively large amount of data observed for each individual.

4.1 Constraining Parameter Variability and Applications to Sequence Clustering

One specific case of the general hierarchical framework results from assuming that there is no parameter variability across individuals within a cluster, or equivalently, that individuals in each group are characterized by the same set of parameters. In the framework of hierarchical models this corresponds to specifying:

$$\pi(\Theta) = \sum_{k=1}^{K} \alpha_k \delta(\Theta - \phi_k),$$

where δ represents the Dirac delta function. This specifies that each individual has the probability α_k of membership in cluster k, but once the cluster membership of an individual is known then parameters are also known to be exactly ϕ_k (i.e., there is no variability within the cluster). With this particular high level model Equation (7) becomes:

$$\begin{aligned} p(Z|D,\Theta,\Phi) &= \prod_{i=1}^{n} \frac{\alpha_{k_i} \delta(\Theta_i - \phi_{k_i}) f(D_i|\Theta_i)}{\sum_{j=1}^{K} \alpha_j \delta(\Theta_i - \phi_j) f(D_i|\Theta_i)} \\ &= \prod_{i=1}^{n} \frac{\alpha_{k_i} f(D_i|\phi_{k_i})}{\sum_{j=1}^{K} \alpha_j f(D_i|\phi_j)} \end{aligned} \tag{16}$$

while Equation (8) becomes:

$$\begin{aligned} w_k^i &= \frac{\alpha_k f(D_i|\phi_k)}{\sum_{j=1}^{K} \alpha_j f(D_i|\phi_j)} \\ Q(\Theta,\Phi;\Theta',\Phi') &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_k^i \log\left[\alpha_k \delta(\Theta_i - \phi_k) f(D_i|\Theta_i)\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_k^i \log\left[\alpha_k f(D_i|\phi_k)\right] \end{aligned} \tag{17}$$

It is interesting to note that this "delta function prior" has implicitly been used in the literature in the past. Krogh et al (1994), and Smyth (1997) used this approach for clustering sequences in an EM framework, using hidden Markov models. Similarly, Smyth (1999) demonstrated the use of EM for clustering individuals based on vector measurements and sequence measurements together and Gaffney and Smyth (1999) uses mixtures of regression models for clustering trajectories and curves. None of these papers cast the problem in a general hierarchical model framework. However, all implicitly assume a delta function prior for the high level model parameters, or equivalently, that there is no within-cluster variation in parameters. The hierarchical model and EM estimation methodology derived in this paper is a strict generalization of this earlier work on probabilistic clustering of non-vector data.

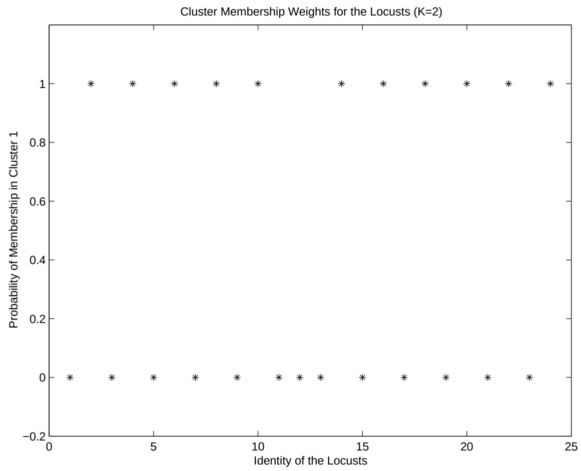

Figure 3: Posterior probabilities of class membership after hierarchical model clustering

We provide a simple example of using a hierarchical model with "delta function priors" to cluster sequence data. The example is intended to be illustrative rather than a systematic investigation of the method. MacDonald and Zucchini (1997) describe behavioral data from 24 locusts in a controlled experiment as shown in Figure 1. Even-numbered locusts were given no food before the experiment and odd-numbered locusts were fed.

We clustered the observed 24 sequences into two groups, assuming no within-cluster variation of parameters across individuals as outlined above. There are $n = 24$ individuals and the observed data D_i for individual i is a binary sequence. We assume a simple first-order Markov model for $f(D_i|\phi_k)$, $k = 1, 2$, for the sequence-generating model in each cluster. Thus, we wish to find a mixture of two Markov models (two clusters of sequential behavior) which account for the data.

The EM equations which result for this specific hierarchical model are quite intuitive and natural. At each EM iteration the likelihood of each sequence D_i is calculated under both high-level models, posterior membership probabilities obtained, and new Markov parameters ϕ_k estimated for each model by weighting the sequences according to their posterior probabilities.

Figure 3 shows the class membership probabilities for cluster 1. This cluster contains the more active (unfed, even-numbered) locusts, with one exception, locust number 12, who although unfed is also relatively inactive, and thus, was clustered with the inactive (fed, odd) locusts. The results demonstrate that the hierarchical model successfully clustered the data into two sensible clusters on this relatively simple problem.

4.2 A Maximum Likelihood Approximation for Large Data Sets, with an Application to Clustering Red Blood Cell Histograms

When there is a large data set D_i available for each individual, the likelihood function associated with the low level model (i.e., the likelihood of parameters Θ_i given the data D_i) will be peaked around its maximum value. If the high level model is reasonably smooth (the prior term in equation (15)), we can approximate Equation (15) by omitting the first term. However, if we omit the first term in equation (15), it becomes a standard ML equation

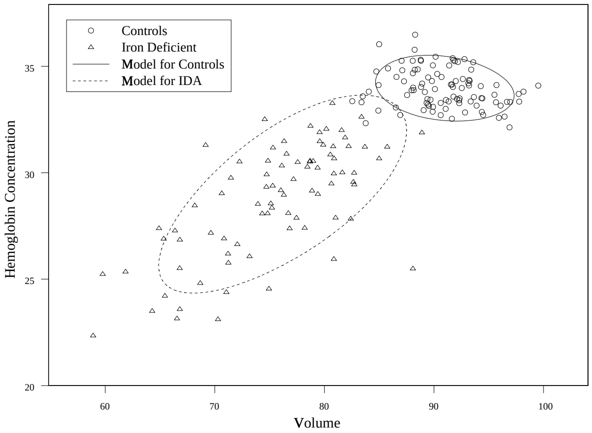

Figure 4: Scatter plot of the estimated mean parameters for individuals from the Iron Deficient and Control groups, with Gaussian "parameter variation models" superposed.

and need not be updated at each iteration. In other words, this approximation *decouples* the high level model and the low level model. It suffices to find the ML estimate $\hat{\Theta}_{iML}$ once for each individual and use the set $\hat{\Theta} = \{\hat{\Theta}_{1ML}, \hat{\Theta}_{2ML}, \ldots, \hat{\Theta}_{nML}\}$ as the data for the ML estimate of Φ at the high level model. This approximation is intuitive: data for each individual are summarized by low level model parameters $\hat{\Theta}_{iML}$, and individuals are then clustered according to the "similarity" of their corresponding parameters, relative to the mixture model prior.

We illustrate the use of this particular approximation with an application in medical diagnosis (Cadez et al, 1999). Blood samples from different individuals are routinely analyzed to screen for various blood-related diseases. An individual blood sample contains around 40,000 individual red blood cells (RBC). Each cell is characterized by its volume (V) and hemoglobin concentration (HC). Thus, for each individual i, D_i is a set of 40,000 such 2-dimensional vectors (one per cell) and we expect the large data set approximation (the decoupled estimation technique described above) to be accurate. It is known (from biological principles) that the V-HC distribution of a single population of cells should be lognormal and that mixtures of log-normals provide a good fit of this type of data. Thus, we use a 2-component bivariate log-normal mixture for the *lower level* model $f(D_i|\Theta_i)$.

The maximum likelihood mixture parameters $\hat{\Theta}_{iML}$ for each of 180 individuals are first estimated. Then the distribution of these parameters is modeled by a mixture prior. Figure 4 shows the fitted mixture model in parameter space for data which is known to come from two classes (normal, or Controls, and iron deficient anemia). The axes in this plot correspond to the estimated bivariate means for each individual, for the larger of the 2 log-normal components in the mixture $f(D_i|\Theta_i)$. The two symbols in the plot correspond to which of the two classes (Control or iron deficient) each patient was diagnosed using a separate accurate laboratory test for iron deficiency. It is important to note that we only use these "ground truth" class labels to illustrate the accuracy of the clustering: the class labels were completely hidden from the hierarchical clustering algorithm. The two ellipses correspond to the covariance matrices of each component $\pi_k(\Theta|\phi_k)$ in the estimated prior, where here we used Gaussian component priors on the means.

This plot is quite informative for a number of reasons. It is clear that the hierarchical model framework matches the observed data quite well here, i.e., there is considerable variability in parameter space, corresponding to the idea of allowing parameters to across individuals. Yet this variability is *systematic* in the sense that the parameter distributions appear to fall into two Gaussian "clouds." Furthermore, these two Gaussian clouds correspond almost exactly to the known prior diagnostic classification of each individual (i.e., most individuals fall into the same cluster as other individuals with the same diagnosis). Since the data sets are so large at the lower-level here, the decoupled and coupled hierarchical models yield virtually identical solutions.

4.3 Systematic Experiments

We ran a number of additional systematic experiments using the red blood cell data described above, where we subsampled both the number of red blood cells per individual (the low level data) and the number of patients. For example, we varied the number of patients in the training data from 10 to 90, and the number of red blood cells per patient from 1,000 to 10,000. In all cases the decoupled approximation and the full hierarchical (coupled) algorithm produced quite similar performance in terms of out-of-sample log-likelihood and classification accuracy (in terms of assigning patients of the same disease to the same clusters). Both models consistently produced accurate clusterings of the data (i.e., similar to that of Figure 4). The results did not provide any significant additional information beyond that shown in Figure 4 and thus, given space constraints, we do not include them here. At the time of writing of the paper we are currently investigating other data sets, applications, and sample size trade-offs, where one might expect the full hierarchical model to provide systematically better results than the decoupled approximation.

5 Conclusions

We presented a general framework for clustering individuals, where we can have measurements on each individual in the form of "non-vector" data (e.g., sequences). Provided the observed data for each individual can be modeled probabilistically, then the hierarchical model provides a relatively realistic model for clustering such data. The key idea is to model variability across different individual's parameters using a mixture model prior on parameters. The component densities in the prior can be directly interpreted as model-based clusters for the individuals. We derived a GEM algorithm for estimating both the prior and the individual parameters. The algorithm is quite tractable and involves an intuitive combination of low-level and high-level estimation steps in the hierarchy. We also derived special cases of the algorithm for problems where there is a large amount of data at the individual level, or where there it is assumed that there is no variability in parameters for each cluster; the latter turns out to be equivalent to a number of existing EM-based probabilistic clustering schemes for clustering sequences and curves. Finally we illustrated the method on two real data sets and demonstrated the utility of the general approach for clustering problems of this nature.

Acknowledgements

The work described in this paper was supported by the National Science Foundation under Grant IRI-9703120. The authors gratefully acknowledge the contributions of Christine

McLaren in providing the red blood cell data.

References

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

Cadez, I.V., McLaren, C. E., Smyth, P., McLachlan, G J., 'Hierarchical models for screening of iron-deficient anemia,' submitted, 1999.

Cheeseman, P. and Stutz. J., 'Bayesian classification (AutoClass): theory and results,' in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), Cambridge, MA: AAAI/MIT Press, pp. 153–180, 1996.

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

Gaffney, S., and Smyth, P., 'Trajectory clustering with mixtures of regression models,' submitted, 1999.

Gelman, A., Carlin, B., Stern, H. S., and Rubin, D., *Bayesian Data Analysis*, London: Chapman and Hall, 1995.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D., 'Hidden Markov models in computational biology: applications to protein modeling,' it J. Mol. Bio., 235:1501–1531, 1994.

Jordan, M. I., and R. A. Jacobs, 'Hierarchical mixtures of experts and the EM algorithm,' *Neural Computation*, 6, 181-214, 1994.

MacDonald, I. and W. Zucchini, *Hidden Markov Models and Other Models for Discrete-Valued Time Series*, Chapman and Hall, 1997.

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

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

Ridgeway, G., 'Finite discrete Markov process clustering,' Technical Report TR 97-24, Microsoft Research, Redmond, WA, 1997.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., and Haussler, D., 'Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology,' *Computer Applications in the Biosciences*, 12:327–345, 1996.

Smyth, P., 'Clustering using Monte-Carlo cross validation,' in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, pp.126–133, 1996.

Smyth, P., ‘Clustering sequences using hidden Markov models,’ in *Advances in Neural Information Processing 9*, M. C. Mozer, M. I. Jordan and T. Petsche (eds.), Cambridge, MA: MIT Press, 648–654, 1997.

Smyth, P., M. Ghil, K. Ide, J. Roden, and A Fraser, ‘Detecting atmospheric regimes using cross-validated clustering,’ *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, 61–66, 1997.

Smyth, P., M. Ghil, and K. Ide, ‘Multiple regimes in Northern hemisphere height fields via mixture model clustering,’ *Journal of Atmospheric Science*, accepted for publication, December 1998.

Smyth, P., ‘Probabilistic model-based clustering of multivariate and sequential data,’ in *Proceedings of the Seventh International Workshop on AI and Statistics*, D. Heckerman and J. Whittaker (eds), San Francisco, CA: Morgan Kaufman.

Thiesson, B., Meek, C., Chickering, D., Heckerman, D., ‘Learning mixtures of DAG models,’ *Proceedings of the Uncertainty in Artificial Intelligence Conference*, San Francisco, CA: Morgan Kaufman, 1998.