Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)

# Partitioning Features for Model-Based Clustering Using Reversible Jump MCMC Technique

**Younghwan Namkoong**
Dept. of CISE
University of Florida
Gainesville, FL 32611, USA
ynamkoon@cise.ufl.edu

**Yongsung Joo**
Dept. of Statistics
Dongguk University
Seoul 100-715, South Korea
yongsungjoo@dongguk.edu

**Douglas D. Dankel II**
Dept. of CISE
University of Florida
Gainesville, FL 32611, USA
ddd@cise.ufl.edu

## Abstract

In many cluster analysis applications, data can be composed of a number of feature subsets where each is represented by a number of diverse mixture model-based clusters. However, in most feature selection algorithms, this kind of cluster structure has been less interesting because they accounted for discovery of a single informative feature subset for clustering. In this study, we attempt to reveal a feature partition comprising multiple feature subsets, with each represented by a mixture model-based cluster. Searching for the desired feature partition is performed by utilizing a local search algorithm based on a reversible jump Markov Chain Monte Carlo technique.

## Introduction

Clustering is a popular data analysis technique and widely used in such application areas as data mining and bioinformatics. To identify interesting patterns in the data, most clustering algorithms deal with all features to represent data. However, only a few features are often relevant to the clustering results, so this requires selecting the proper subset of features to represent the clusters. To improve the performance of clustering algorithms, this feature selection technique has been an interesting problem in spite of the absence of prior information (Sayes, Inza, and Larrañaga 2007). Previous feature selection approaches that have investigated efficient techniques to select relevant features for clustering commonly assume that features are divided into two feature subsets (Constantinopoulos, Titsias, and Likas 2006). However, the original feature vectors can consist of a number of feature subsets, making these previous feature selection approaches unsuitable.

In this paper, we propose a novel approach to find a set of the feature subsets based on the gaussian mixture model at the same time. For each feature subset, the best-fit mixture model to represent clusters is determined by maximizing the entropy of the maximum likelihood estimates, achieved via the deterministic annealing expectation maximization (DAEM) algorithm. By collecting these feature subsets, the desired feature partition is obtained. To avoid an unreasonable processing time when searching for the desired feature partition, the simulated annealing-based reversible jump Markov Chain Monte Carlo technique (RJMCMC) technique is utilized. Through some numerical examples, we show that our approach is insensitive to the various initial feature partitions.

## Mixture model-based clustering on the partitioned feature subsets

Let $\mathbf{X}$ be a data matrix with $N$ data objects, each of which is represented by a $D$-dimensional feature vectors, $\mathbf{v} = (v_1, \ldots, v_d, \ldots, v_D)$. Given $\mathbf{v}$, a partition of features $\mathcal{V}$ consists of $K$ disjoint nonempty subsets of features, denoted by $\mathcal{V} = \{\mathcal{V}_k;\ k \in \{1, \ldots, K\}$ and $K \in \{1, \ldots, D\}\}$, where $\mathcal{V}_k$ is the $k^{th}$ subset of features. In particular, $\bigcup_{k=1}^{K} \mathcal{V}_k = \mathbf{v}$ and $\mathcal{V}_k \cap \mathcal{V}_{k'} = \emptyset$, for any $k' \in \{1, \ldots, K\}$ and $k \neq k'$. $U_{\mathcal{V}_k}$ is a submatrix of $\mathbf{X}$ corresponding to $\mathcal{V}_k$. Based on the concept of mixture model-based clustering, our approach assumes that all $\mathcal{V}_k$s are mutually independent and each $U_{\mathcal{V}_k}$ lies in different gaussian mixture models with different mixture components. Then, the log-likelihood function of $\mathbf{X}$, $\mathcal{L}(\theta|\mathbf{X})$, can be expressed as $\mathcal{L}(\theta|\mathbf{X}) = \sum_{k=1}^{K} \mathcal{L}_k(\theta_k|U_{\mathcal{V}_k})$. The estimates maximizing $\mathcal{L}_k(\theta_k|U_{\mathcal{V}_k})$ for fixed $\mathcal{V}_k$ and $G_k$ can be usually achieved via the EM algorithm. For $\mathcal{V}_k$, let $Z$ be missing variables, where $Z_{kgn} = 1$ or 0 if the $n^{th}$ object is assigned to $C_{\mathcal{V}_k g}$ or not. Then, the complete data log-likelihood function is

$$\mathcal{L}_k(\theta_k|U_{\mathcal{V}_k}, Z_k) = \sum_{n=1}^{N} \log \sum_{g=1}^{G_k} Z_{kgn} p_{kg} \phi(U_{\mathcal{V}_k n}; \theta_{kg}). \quad (1)$$

where $\theta_k = (\theta_{k1}, \ldots, \theta_{kg}, \ldots, \theta_{kG_k})$. $\theta_{kg}$ and $\phi(U_{\mathcal{V}_k n}|\theta_{kg})$ are the parameter values and the gaussian probability density function of the $g^{th}$ cluster, respectively.

Starting with an initial parameter values $\theta_k^{(0)}$, the EM algorithm alternates the E-step and the M-step to update $\theta_k$. In the $i^{th}$ E-step, the conditional expectation of the complete data log-likelihood, called the $Q$ function, is computed:

$$Q(\theta_k, \theta_k^{(i)}) = E[\mathcal{L}_k(\theta_k|U_{\mathcal{V}_k}, Z_k)|U_{\mathcal{V}_k}; \theta_k^{(i)}]. \quad (2)$$

In the M-step, new parameter estimates, $\theta_k^{(i+1)}$, maximizing $Q(\theta_k, \theta_k^{(i)})$ are calculated. This process stops when a convergence condition is satisfied (Theodoridis and Koutroumbas 2006).

To mitigate the local maxima problem of the EM algorithm due to the monotonic convergence property, one can consider utilizing the deterministic annealing EM (DAEM) algorithm that uses a modified log-likelihood including the "thermodynamic free energy" parameter $\beta$ ($0 < \beta < 1$) (Theodoridis and Koutroumbas 2006). Specifically, the DAEM algorithm starts with a small initial $\beta$, which is close to 0. Then, until $\beta$ becomes 1, the DAEM algorithm performs the E and M steps by gradually increasing $\beta$ to obtain a better local (and possibly global) maximum.

The estimated model can vary depending on the values of the parameters, so an appropriate model amongst many candidate models should be selected. One natural way for this problem is to choose a model which is the most similar to the "true" model. For measuring the similarity between the true model and the estimated model, it is a good choice to utilize the Akaike Information Criterion (AIC), a well-known model selection method based on estimating the Kullback-Leibler (KL) divergence. In the model selection process, an estimated model is regarded as the best fitted model when the score of AIC is minimized. For the given $\mathcal{V}_k$, $AIC(\mathcal{V}_k)$ is denoted by

$$AIC(\mathcal{V}_k) = -2 \times \mathcal{L}_k(\hat{\theta}_k|U_k) + 2\lambda_k, \tag{3}$$

where $\mathcal{L}_k(\hat{\theta}_k|U_k)$ is the maximum log-likelihood, and $\lambda_k$ is the number of parameters $\hat{\theta}_k$. By aggregating $AIC(\mathcal{V}_k)$s, feature partition consisting of multiple feature blocks where each can be expressed by the best-fit mixture model for clustering can be obtained. Accordingly, the AIC for $\mathcal{V}$, used for an objective function in our approach, can be expressed by:

$$J(X, \mathcal{V}, \theta) = \sum_{k=1}^{K} AIC(\mathcal{V}_k). \tag{4}$$

Searching for the feature partition minimizing the objective function (4) through exhaustive search is quite challenging because the number of all possible partitions for a fixed number of features grows hyper-exponentially. Moreover, for each feature subset $\mathcal{V}_k$ in a given feature partition $\mathcal{V}$, the model selection process as well as the mixture model-based clustering via the DAEM algorithm aggravate the prohibitive computational complexity.

For such a combinatorial optimization problem, we attempt to search the desired feature partition $\mathcal{V}^*$ by using a local search algorithm, called the biased random walk algorithm (Booth, Casella, and Hobert 2008). However, this algorithm has several drawbacks such as the relative narrow search region at each state and/or a local maxima problem. This problem can be overcome by embedding another annealing process, called the Simulated Annealing (SA) technique, with the biased random walk algorithm. In SA, the temperature parameter $T$ ($T > 0$) represents the degree of random transition between states, meaning that a candidate state tends to be accepted at a high temperature. Our search algorithm starts with the initial feature partition $\mathcal{V}^{(0)}$ and the initial temperature parameter $T^{(0)}$ set to a high value. Until the final state becomes stable, when $T \rightarrow 0$, this search algorithm explores $\mathcal{V}^*$ by gradually decreasing $T$. At the $t^{th}$ state, a candidate feature partition $\mathcal{V}'^{(t)}$ is accepted by the following probability $\gamma'$:

$$\gamma' = \min\left[1, \alpha = \frac{\pi(\mathcal{V}'^{(t)})}{\pi(\mathcal{V}^{(t)})}\exp\{T^{(t)}\}\right], \tag{5}$$

where $T^{(t)} = \rho \times T^{(t-1)}$, $t \geq 1$, and $\rho$ is a cooling rate, $\rho \in (0, 1)$. $\pi(\mathcal{V})$ be a probability mass function related with $\mathcal{V}$.

## Simulation and Discussion

Three synthetic datasets for experiments were generated on the basis of the following parameters: $K$, $G_k$s, $\mathcal{V}$, $p_{kg}$s, $\mu_{gd}$s, and $\Sigma_{gd}$s where $g \in \{1, \ldots, G_k\}$, $d \in \{1, \ldots, D\}$, and $k \in \{1, \ldots, K\}$. For $\mathcal{V}$, $\mathcal{V}_k$ is composed of a different number of $\mathcal{C}_{\mathcal{V}_k g}$ and its corresponding $p_{kg}$. Each $X_n$ lies in a gaussian distribution corresponding to $\mu_{kg}$ and $\Sigma_{kg}$. For example, the 1st dataset contains three feature subsets: $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$. Each $\mathcal{V}_k$ has 3, 2, and 1 mixture components, respectively. The other datasets have different shapes and more complex structures than the 1st dataset. In particular, the 2nd dataset has a checkerboard structure.

To cover the overall cases for each dataset, we used 5 different initial partition $\mathcal{V}^{(0)}$s. Specifically, each $\mathcal{V}_k$ in 1) is a singleton feature subset, $\mathcal{V}^{(0)}$ of 2) is a feature subset with all features, and the remaining three $\mathcal{V}^{(0)}$s were randomly generated. To support enough randomness, $T^{(0)} = 400.0$ and $\rho = 0.997$. For the DAEM algorithm, the $\beta^{(0)} = 4.0$ and $\beta^{(i)} = \beta^{(i-1)} \times 0.998$. Through the simulation results, our search method demonstrated insensitivity to the various initial feature partitions by showing successful convergence to the minimum score of the objective function and the parameter estimates near the true parameter values.

Our method can be useful in various application areas. For example, in Bioinformatics, gene expression data consisting of genes (column) and experimental conditions (row) can be expressed by multiple gene groups where each corresponds to a number of different condition clusters across the above diverse gene subsets. Relevant experiments in various application areas are currently in progress.

## References

Booth, J. G.; Casella, G.; and Hobert, J. P. 2008. Clustering using objective functions and stochastic search. *Journal Of The Royal Statistical Society Series B* 70(1):119–139.

Constantinopoulos, C.; Titsias, M. K.; and Likas, A. 2006. Bayesian feature and model selection for gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(6):1013–1018.

Sayes, Y.; Inza, I.; and Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. *BIOINFORMATICS* 23(9):2507–2517.

Theodoridis, S., and Koutroumbas, K. 2006. *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc.