A Gaussian Mixture Model to Detect Clusters Embedded in Feature Subspace

Yuanhong Li, Ming Dong and Jing Hua
Department of Computer Science
Wayne State University, Detroit, MI 48202

Abstract

The goal of unsupervised learning, i.e., clustering, is to determine the intrinsic structure of unlabeled data. Feature selection for clustering improves the performance of grouping by removing irrelevant features. Typical feature selection algorithms select a common feature subset for all the clusters. Consequently, clusters embedded in different feature subspaces are not able to be identified. In this paper, we introduce a probabilistic model based on Gaussian mixture to solve this problem. Particularly, the feature relevance for an individual cluster is treated as a probability, which is represented by localized feature saliency and estimated through Expectation Maximization (EM) algorithm during the clustering process. In addition, the number of clusters is determined simultaneously by integrating a Minimum Message Length (MML) criterion. Experiments carried on both synthetic and real-world datasets illustrate the performance of the proposed approach in finding clusters embedded in feature subspace.

1 Introduction

Clustering is unsupervised classification of data objects into different groups (clusters) such that objects in one group are similar together and dissimilar from another group. Applications of data clustering are found in many fields, such as information discovering, text mining, web analysis, image grouping, medical diagnosis, and bioinformatics. Many clustering algorithms have been proposed in the literature [8]. Basically, they can be categorized into two groups: hierarchical or partitional. A clustering algorithm typically considers all available features of the dataset in an attempt to learn as much as possible from data. In practice, however, some features can be irrelevant, and thus hinder the clustering performance. *Feature selection*, which chooses the "best" feature subset for clustering, can be applied to solve this problem.

Feature selection is extensively studied in supervised learning scenario [1–3], where class labels are available for judging the performance improvement contributed by a feature selection algorithm. For unsupervised learning, feature selection is a very difficult problem due to the lack of class labels, and it has received extensive attention recently. The algorithm proposed in [4] measures feature similarity by an information compression index. In [5], the relevant features are detected using a distance-based entropy measure. [6] evaluates the cluster quality over different feature

subsets by normalizing cluster separability or likelihood using a cross-projection method. In [7], feature saliency is defined as a probability and estimated by the Expectation Maximization (EM) algorithm using Gaussian mixture models. A variational Bayesian approach is presented in [9]. The algorithm described in [10] employs a criterion on the psychological similarity for content-based image retrieval systems. An evolutionary local selection algorithm is used in [11] to search for possible combination of features and numbers of clusters, with the guidance of the k-means algorithm. The benefits of feature selection include simplifying the problem by discarding irrelevant information, improving the learning performance, reducing the storage cost of databases, and providing more precise knowledge of the underlying model that generates the data.

The aforementioned algorithms perform feature selection in a *global* sense by producing a common feature subset for all the clusters. This, however, can be problematic in practice, where the local intrinsic property of data matters more for grouping analysis [12]. In the illustrative example shown in Figure 1, the relevant feature subset for cluster C_1 is $\{x_1, x_2\}$, while clusters C_2 and C_3 are better to be recognized on $\{x_2\}$ and $\{x_1\}$, respectively. A common feature subset, i.e., $\{x_1, x_2\}$, can not reflect the inherent structural properties of the three clusters. Clustering with local features is highly desired. To this end, bipartite graph partitioning algorithms [13, 14] attempt to partition features together with patterns such that the output contains relevant features for each individual cluster. However, features are divided exclusively, which prevents a feature to be relevant to more than one cluster. Other approaches in this direction, usually referred as *subspace clustering* [15], seek density areas embedded in a high dimensional feature space [16–20]. These algorithms navigate the possible subspaces heuristically [20] or in a grid manner [16], often requiring the density threshold and the cluster number as inputs. In addition, the clusters produced are overlapping in many cases.

In this paper, we focus on the clustering problems with exclusive partitioning. We propose to detect clusters embedded in feature subspace based on EM with a local feature saliency measure. The number of clusters is also simultaneously detected by integrating a Minimum Message Length (MML) criterion. Through experiments performed on both synthetic and real-world datasets, we demonstrate the advantages of the proposed localized feature selection method over the global one. The rest of the paper is organized as follows: In Section 2, we introduce some essential background on EM-based clustering and simultaneous global feature selection. In Section 3, we perform model detection for Gaussian mixture through EM with an integrated local feature saliency. The proposed algorithm is evaluated on both synthetic and real-world datasets in Section 4. Finally, we summarize our work in Section 5.

2 Background on EM-based Clustering and Global Feature Selection

From a *model-based* perspective, each cluster can be mathematically represented by a parametric distribution. The entire dataset is therefore modeled by a mixture of these distributions. The most widely used model in practice is

the mixture of Gaussians. The clustering process thereby turns to estimating the parameters of the Gaussian mixture, usually by the EM algorithm.

Traditionally, a finite mixture of densities with K components is represented by,

$$p(y) = \sum_{j=1}^{K} \alpha_j p(y|\theta_j), \tag{1}$$

where α_j is the *a priori* probability, and θ_j is a set of parameters of component j. The parameters are estimated by maximizing the likelihood as,

$$\hat{\theta}_{ML} = \arg\max_{\theta} \left[\log p(\mathcal{Y}|\theta)\right]. \tag{2}$$

Let $\mathcal{Z} = \{z_{ij}\}_{N\times K}$ be a set of missing (latent) cluster labels, where $z_{ij} = 1$ if y_i is a sample of $p(\cdot|\theta_j)$, and $z_{ij} = 0$ otherwise. $\mathcal{Z}$ can be also written as a vector $\mathcal{Z} = (z_1, \ldots, z_N)$ such that $z_i = j$ if y_i is a sample of $p(\cdot|\theta_j)$. The log-likelihood when $\mathcal{Z}$ is observed is,

$$\log p(\mathcal{Y}, \mathcal{Z}|\theta) = \sum_{i=1}^{N}\sum_{j=1}^{K} z_{ij} \log[\alpha_j p(y_i|\theta_j)] \tag{3}$$

Let $\mathcal{W} = \mathrm{E}[\mathcal{Z}|\mathcal{Y}, \hat{\theta}(t)]$ represent the expected value of $\mathcal{Z}$, where $\hat{\theta}(t)$ is the estimate of θ at iteration t. The parameters can be estimated by the following updating rule,

$$\hat{\theta}(t+1) = \arg\max_{\theta}\{\log p(\mathcal{Y}, \mathcal{W}|\hat{\theta}(t))\} \tag{4}$$

Assuming features are conditionally independent, the mixture of densities can be described as,

$$p(y|\theta) = \sum_{j=1}^{K} \alpha_j p(y|\theta_j) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} p(y_l|\theta_{jl}) \tag{5}$$

where D is the number of features. Define the global feature saliency ρ_l to be the probability that feature l is salient to all the components. Then, $(1 - \rho_l)$ is the probability that l is not salient to any of the components. Let $\Phi = (\phi_1, \ldots, \phi_D)$ be the feature relevance vector with $\phi_l = 1$, if feature l is relevant and, $\phi_l = 0$, otherwise. Then, $\rho_l = \Pr(\phi_l = 1)$. Finally, the likelihood function can be rewritten as [7],

$$p(y|\theta) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} [\rho_l p(y_l|\theta_{jl}) + (1 - \rho_l) q(y_l|\lambda_l)] \tag{6}$$

where $q(\cdot(\lambda_l))$ is a common density, which defines the distribution of an irrelevant feature l. If we treat Φ as missing variables, the feature saliency vector ρ can be estimated by the EM algorithm [7].

3 Detecting Clusters Embedded in Feature Subspace

In this section, we present a probabilistic model based on Gaussian mixture to detect clusters embedded in feature subspace. First, we define a localized feature saliency and show how it could be integrated into EM clustering. Then, we estimate the number of clusters with the MML criterion.

3.1 Localized Feature Saliency

In our approach, the importance of a feature can be different for different clusters, which implies that the feature relevance takes a matrix form, $\Phi = \{\phi_{jl}\}_{K\times D}$, where $\phi_{jl} = 1$ indicates that feature l is associated with component j, otherwise $\phi_{jl} = 0$. Let $\rho_{jl} = \Pr(\phi_{jl} = 1)$ be the probability that feature l is relevant to component j. Then, the likelihood can be obtained based on the following proposition.

Proposition 1. *Let $p(\cdot|\theta_{jl})$ represent the distribution of a salient feature l for a particular component j, and $q(\cdot|\lambda_{jl})$ the distribution if feature l is non-salient to the particular component. Assuming that the features are conditionally independent, the likelihood function can be written as,*

$$p(y|\theta) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} \big(\rho_{jl} p(y_l|\theta_{jl}) + (1-\rho_{jl}) q(y_l|\lambda_{jl})\big) \tag{7}$$

Proof. Let $\phi_j = (\phi_{j1}, \ldots, \phi_{jD})$. For a particular component j, we have

$$\begin{aligned}
p(y|z=j,\phi_j) &= \prod_{l=1}^{D} \big(p(y_l|\theta_{jl})\big)^{\phi_{jl}} \big(q(y_l|\lambda_{jl})\big)^{1-\phi_{jl}} \\
p(y,\phi_j,z=j) &= p(y|z=j,\phi_j)p(\phi_j|z=j)P(z=j) \\
&= \alpha_j \prod_{l=1}^{D} \big(p(y_l|\theta_{jl})\big)^{\phi_{jl}} \big(q(y_l|\lambda_{jl})\big)^{1-\phi_{jl}} \prod_{l=1}^{D} \rho_{jl}^{\phi_{jl}} (1-\rho_{jl})^{1-\phi_{jl}} \\
&= \alpha_j \prod_{l=1}^{D} \big(\rho_{jl} p(y_l|\theta_{jl})\big)^{\phi_{jl}} \big((1-\rho_{jl}) q(y_l|\lambda_{jl})\big)^{1-\phi_{jl}}
\end{aligned} \tag{8}$$

Marginal density on y gives

$$\begin{aligned}
p(y|\theta) &= \sum_{j,\Phi}^{K} p(y,\phi_j,z=j) \\
&= \sum_{j=1}^{K} \alpha_j \sum_{\phi_{jl}} \prod_{l=1}^{D} \left(\rho_{jl} p(y_l|\theta_{jl})\right)^{\phi_{jl}} \left((1-\rho_{jl}) q(y_l|\lambda_{jl})\right)^{1-\phi_{jl}} \\
&= \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} \left(\rho_{jl} p(y_l|\theta_{jl}) + (1-\rho_{jl}) q(y_l|\lambda_{jl})\right)
\end{aligned} \tag{9}$$

where $\theta = \{\{\alpha_j\}, \{\theta_{jl}\}, \{\rho_{jl}\}, \{\lambda_{jl}\}\}$ is the set of all the parameters. □

Taking $\{z_{ij}\}$ and $\{\phi_{jl}\}$ as latent variables, we derive the E-step and M-step of the EM algorithm to estimate the parameter set.

E-Step: Compute the expectation of the log-likelihood.

From Equation (8), the expected complete log-likelihood of the dataset based on $\theta^{(t)}$ is

$$
\begin{aligned}
& E_{\theta^{(t)}}[\log P(\mathcal{Y}, z, \Phi)] \\
=& \sum_{i,j,\Phi} P(z_i = j, \Phi | y_i)(\log \alpha_j + \sum_l \phi_{jl} \left(\log \rho_{jl} + \log p(y_{il}|\theta_{jl})\right) \\
& + (1 - \phi_{jl}) \left(\log(1 - \rho_{jl}) + \log q(y_{il}|\lambda_{jl})\right)) \\
=& \sum_j (\sum_i P(z_i = j | y_i)) \log \alpha_j \\
& + \sum_{jl} \sum_i P(z_i = j, \phi_{jl} = 1 | y_i)(\log p(y_{il}|\theta_{jl}) + \log \rho_{jl}) \\
& + \sum_{jl} \sum_i P(z_i = j, \phi_{jl} = 0 | y_i) \left(\log q(y_{il}|\lambda_{jl}) + \log(1 - \rho_{jl})\right)
\end{aligned} \tag{10}
$$

The probabilities are computed as follows,

$$
P(z_i = j | y_i) = \frac{\alpha_j \prod_l [\rho_{jl} p(y_{jl}|\theta_{jl}) + (1 - \rho_{jl}) q(y_{jl}|\lambda_{jl})]}{\sum_j \alpha_j \prod_l [\rho_{jl} p(y_{jl}|\theta_{jl}) + (1 - \rho_{jl}) q(y_{jl}|\lambda_{jl})]} \tag{11}
$$

$$
P(z_i = j, \phi_{jl} = 1 | y_i) = \frac{\rho_{jl} p(y_{jl}|\theta_{jl})}{\rho_{jl} p(y_{jl}|\theta_{jl}) + (1 - \rho_{jl}) q(y_{jl}|\lambda_{jl})} P(z_i = j | y_i) \tag{12}
$$

$$
P(z_i = j, \phi_{jl} = 0 | y_i) = \frac{(1 - \rho_{jl}) q(y_{jl}|\lambda_{jl})}{\rho_{jl} p(y_{jl}|\theta_{jl}) + (1 - \rho_{jl}) q(y_{jl}|\lambda_{jl})} P(z_i = j | y_i) \tag{13}
$$

M-step: Maximize the log-likelihood.

The three parts of Equation (10) can be maximized separately by updating the following quantities,

$$\widehat{\alpha_j} = \frac{\sum_i P(z_i = j|y_i)}{\sum_j \sum_i P(z_i = j|y_i)} \tag{14}$$

$$\widehat{\mu_{\theta_{jl}}} = \frac{\sum_i P(z_i = j, \phi_{jl} = 1|y_i) y_{jl}}{\sum_i P(z_i = j, \phi_{jl} = 1|y_i)} \tag{15}$$

$$\widehat{\sigma^2_{\theta_{jl}}} = \frac{\sum_i P(z_i = j, \phi_{jl} = 1|y_i)(y_{jl} - \widehat{\mu_{\theta_{jl}}})^2}{\sum_i P(z_i = j, \phi_{jl} = 1|y_i)} \tag{16}$$

$$\widehat{\mu_{\lambda_{jl}}} = \frac{\sum_i P(z_i = j, \phi_{jl} = 0|y_i) y_{jl}}{\sum_i P(z_i = j, \phi_{jl} = 0|y_i)} \tag{17}$$

$$\widehat{\sigma^2_{\lambda_{jl}}} = \frac{\sum_i P(z_i = j, \phi_{jl} = 0|y_i)(y_{jl} - \widehat{\mu_{\lambda_{jl}}})^2}{\sum_i P(z_i = j, \phi_{jl} = 0|y_i)} \tag{18}$$

$$\widehat{\rho_{jl}} = \frac{\sum_i P(z_i = j, \phi_{jl} = 1|y_i)}{\sum_i P(z_i = j, \phi_{jl} = 1|y_i) + \sum_i P(z_i = j, \phi_{jl} = 0|y_i)} \tag{19}$$

The EM algorithm alternates between the E-step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and the M-step, which maximizes the expected likelihood found in the E-step. The parameters found in the M-step are then used to begin another iteration of the E-step, and the process is continued until the algorithm converges to a finite mixture model with feature saliency associated with each cluster. Thus, clustering and localized feature saliency detection is achieved simultaneously.

3.2 Model Selection Based on Minimum Message Length (MML)

Alternation of E and M steps in the above algorithm eventually results in a maximum likelihood estimate of Gaussian mixtures, which requires the number of clusters K as prior knowledge. To overcome this difficulty, we employ the MML criterion to detect the optimal number of clusters [7]. The MML criterion for our model with respect to θ is as follows,

$$\begin{aligned} J(\theta) = &- \log(\mathcal{Y}|\theta) + \frac{1}{2}(K + DK)\log(N) \\ &+ \frac{R}{2}\sum_{l=1}^{D}\sum_{j=1}^{K}\log(N\alpha_j\rho_{jl}) + \frac{S}{2}\sum_{l=1}^{D}\sum_{j=1}^{K}\log(N\alpha_j(1-\rho_{jl})) \end{aligned} \tag{20}$$

In the above equation, R and S are the number of parameters of $p(\cdot)$ and $q(\cdot)$, respectively, which for a Gaussian distribution is 2. Also, $-\log(\mathcal{Y}|\theta)$ corresponds to log-likelihood, and $\frac{1}{2}(K + DK)\log(N)$ represents the code-length of standard Message Description Length (MDL) of parameters α_js and ρ_{jl}s. While $N\alpha_j\rho_{jl}$ indicates the effective number of data for estimating θ_{jl}, $\frac{R}{2}\sum_{l=1}^{D}\sum_{j=1}^{K}\log(N\alpha_j\rho_{jl})$ computes the code-length corresponding to the parameters θ_{jl}. Similarly, $\frac{S}{2}\sum_{l=1}^{D}\sum_{j=1}^{K}\log(N\alpha_j(1-\rho_{jl}))$ represents the code-length for parameters λ_{jl}. The

optimal mixture model is the one that minimizes the cost function $J(\theta)$ in Equation (20),

$$\hat{\theta} = \arg\min_{\theta}(J(\theta)) \tag{21}$$

The algorithm introduced above works well in general cases. However, extreme bad initialization may lead to some clusters with singular covariance matrices, and thus adversely affect the cost function $J(\theta)$. Those clusters can be pruned based on a modification of Equation (14) [7],

$$\widehat{\alpha_j} = \frac{\max\left(\sum_i P(z_i = j|y_i) - \frac{RD}{2}, 0\right)}{\sum_j \max\left(\sum_i P(z_i = j|y_i) - \frac{RD}{2}, 0\right)} \tag{22}$$

The effect of Equation (22) is that some small trivial components are quickly eliminated at an early stage. Similarly, Equation (19) is modified to,

$$\widehat{\rho_{jl}} = \frac{\max\left(\sum_i P(z_i = j, \phi_{jl} = 1|y_i) - \frac{R}{2}, 0\right)}{\max\left(\sum_i P(z_i = j, \phi_{jl} = 1|y_i) - \frac{R}{2}, 0\right) + \max\left(\sum_i P(z_i = j, \phi_{jl} = 0|y_i) - \frac{S}{2}, 0\right)} \tag{23}$$

The above Equation can prune ρ_{jl} to either 1 or 0.

In summary, the proposed EM clustering with localized feature saliency consists of the following steps,

1. Initialize the algorithm with a large value of K, minimal number of components K_{min}, and the parameter set θ.
2. Alternate between E-step and M-step until the model converges to a local maximum. During this step, components with $\alpha_j = 0$ are pruned.
3. Record the parameter set θ and the message length based on Equation (20).
4. Terminate the iterations if K equals K_{min}. Otherwise, reduce K to $K-1$ by removing the smallest component, and repeat steps (2) and (3).
5. Output the model with the smallest message length.

3.3 Computational Complexity

The computational load of the proposed algorithm is mainly due to the E and M steps. For every iteration, the complexity of both the steps is $\mathcal{O}(KND)$. The total computational time is dependent on the number of iterations required for converging. Conventional feature selection algorithms usually seek optimal features by trying out large number of combinations. On the other hand, the proposed algorithm computes the localized feature saliency simultaneously with clustering, thus avoiding the navigation over all possible feature subsets. It only needs to search over a small set of possible Ks.

4 Experimental Results

In general, the performance of an unsupervised feature selection algorithm is hard to be evaluated. Localized feature selection makes it even more difficult as we have an additional layer of complexity brought by the association of clusters to different feature subsets. In this section, we provide thorough evaluation of the proposed algorithm by comparing it with the global feature selection approach [7] on both synthetic and real-world datasets. In addition, we show the need for feature selection in clustering and the benefits of selecting features locally through a case-study on Boston housing dataset.

4.1 Synthetic Data

First, we applied both our method and the global feature selection algorithm to several synthetic datasets. As we know the underlying models from which the patterns were sampled from, the performance of an algorithm is interpreted as: can the algorithm find the given model? The synthetic datasets are created by a data generator. It first generates c Gaussian components $\mathcal{N}(\mu_j, \Sigma_j)$, $j = 1, \cdots, c$, separately, where Σ_j is restricted to a diagonal matrix. Components can have different number of features D_j, and different number of patterns N_j. Those Gaussians are then embedded into subsets of a D-dimensional background with Gaussian noise $\mathcal{N}(0, I)$. Finally, a D-dimensional dataset consisting of c Gaussian mixtures, with each component corresponding to an individual relevant feature subset is generated. The total number of patterns is $N = \sum_{j=1}^{c} N_j$. Table 1 shows a summary of the four synthetic datasets generated.

In the experiments, we initialized the parameters as follows: number of clusters K is set to 20, the *a priori* probabilities α_j are set equally at $1/20$, the feature saliencies ρ_{jl} are set at 0.5, and the common components are set to cover the entire dataset. We ran the proposed algorithm 10 times independently with stopping threshold of 10^{-7}. The clustering error rates and cluster numbers are computed as the average over the 10 runs, and standard deviations are calculated accordingly. The feature saliency for each cluster at each run is mapped to a grey-scale image, where each column represents a feature, and each row represents an individual run, as shown in Table 2. For all the four datasets, the proposed algorithm successfully detected the number of clusters. Each cluster and its relevant feature subset are also detected correctly. The grey-scale image is steady vertically, indicating that the algorithm is stable in different runs. In Table 2, we also show the performance of the global feature selection algorithm [7] on each of the datasets. We can see that the union of the localized feature subsets is equivalent to the relevant features selected by the global approach. Moreover, while global algorithm is able to detect the number of clusters correctly, it can not determine if a salient feature really plays a critical role for a particular cluster. On the other hand, our approach yields more informative models, which not only provide information about whether a feature is relevant or not, but also about which cluster the feature is relevant or irrelevant to.

4.2 Real-world datasets

For the evaluation on real-world datasets, we utilized four datasets: *wine*, *wdbc*, *vehicle*, and *zernike*, from the UCI machine learning repository [21], having varying number of features, patterns, and categories. The *wine* dataset is used to recognize different wine types by 13 characters of chemical analysis. It consists of 178 patterns and 3 categories. The *wdbc* dataset is used to diagnose if a breast cancer is benign or malignant based on 30 features and contains 576 data points. The *vehicle* dataset contains 846 samples with 18 features extracted from vehicle silhouettes. The purpose is to classify a given silhouette as one of four types of vehicles. The *zernike* dataset records 47 zernike moments extracted from 2000 images of handwriting digits. Summary of these four datasets is shown in Table 3. The parameters are initialized in the same way as for the synthetic datasets, except that K is set at 30 for the *zernike* dataset.

The datasets are provided with class labels for supervised learning, which are excluded during the clustering process. We assign a class label to each final cluster afterwards so that a pseudo error rate can be computed for evaluation purpose. The cluster label is simply selected as the class to which majority of patterns in the cluster belongs. In other words, we assume that each cluster consists of patterns from the same class. Comparing the cluster labels of all the patterns with the true class labels yields the pseudo error rate.

The estimated cluster numbers and pseudo error rates are shown in Table 4 for both local and global methods. It is clear that the proposed EM clustering with localized feature saliency generally outperforms the global one with lower error rates and variances. We also compared the feature saliency of the two algorithms as grey-scale images in Table 5. Obviously, different clusters have different relevant feature subsets, which are usually smaller than the globally relevant feature subset. This result indicates that a globally relevant feature can be irrelevant to some clusters. Our experiments also show that a locally relevant feature might be treated as globally irrelevant. For example, the third feature of *wine* dataset is relevant to the first cluster (bright column), but, it has been ignored by the global feature selection algorithm (dark column). Thus, EM clustering with localized feature saliency provides users more accurate knowledge regarding the underlying model from which the cluster component is generated. Moreover, the vertical belt patterns in the grey-scale images demonstrates the stability of the proposed algorithm over different runs.

4.3 Boston Housing Dataset

In this section, we present a case study of the proposed algorithm on the Boston housing data from UCI [21], which contains 506 neighborhoods in the Boston metropolitan area with 14 attributes, as described in Table 6. This dataset is often used as a test bed to compare the performance of prediction methods by estimating the value of the last attribute MEDV from the other 13 attributes. In our experiment, we remove the binary attribute CHAS, and consider the rest 13 attributes on an equal basis. Our goal is to find groups of neighborhoods based on these attributes, and to identify

the saliency of attributes for each individual group.

In our experiment, the number of clusters are initialized to 20, and other parameters are initialized in the same way as for the synthetic datasets. As shown in Figure 2, 10 clusters are identified. Notice that the attribute saliency varies for each cluster. For example, attributes {CRIM, RAD, TAX, PTRT} are important to Group A but not to Group E, while attribute B is important to Group E but not to Group A. Figure 2 clearly shows that the distribution of feature saliency over the 13 attributes is quite different across clusters. Traditional clustering algorithms without feature selection or with global feature selection is not able to reveal these properties of the dataset. Our method, on the other hand, can provide this vital information to users through cluster-wise feature selection.

5 Conclusion

In this paper, we proposed a EM clustering algorithm with localized feature saliency. In our approach, unsupervised feature selection is performed by estimating feature saliency of individual clusters simultaneously with the EM clustering. The determination of cluster number is also integrated in our method by adopting an MML criterion. Experimental results show that the cluster model produced by the proposed algorithm can provide users more accurate understanding of the underlying process which generates the data.

Acknowledgement

This research was partially funded by the 21st Century Jobs Fund Award, State of Michigan, under grant: 06-1-P1-0193, and by National Science Foundation, under grant: IIS-0713315.

References

[1] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[2] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[3] M. Dong and R. Kothari, "Feature subset selection using a new definition of classifiability," *Pattern Recognition Letters*, vol. 23, pp. 1215–1225, 2003.

[4] P. Mitra, C. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 301 –312, 2002.

[5] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering - a filter solution," in *IEEE International Conference on Data Mining*, 2002, pp. 115–122.

[6] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning." *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[7] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.

[8] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM Coimputing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[9] C. Constantinopoulos, M. K. Titsias, and A. Likas, "Bayesian feature and model selection for gaussian mixture models," *IEEE Trans. on PAMI*, vol. 28, no. 6, pp. 1013–1018, 2006.

[10] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based imageretrieval," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 702–712, 2006.

[11] Y. S. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 365–369.

[12] Y. Li, M. Dong and J. Hua, "Localized feature selection for clustering," *Pattern Recognition Letters*, in press, doi:10.1016/j.patrec.2007.08.012.

[13] H. Zha, X. He, C. Ding, M. Gu, and H. Simon, "Bipartite graph partitioning and data clustering," in *Proceedings of ACM CIKM*, 2001, pp. 25–32.

[14] M. Rege, M. Dong, and F. Fotouhi, "Co-clustering documents and words using bipartite isoperimetric graph partitioning," in *IEEE International Conference on Data Mining (ICDM)*, Hong Kong, 2006, pp. 532 – 541.

[15] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, 2004.

[16] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *1998 ACM-SIGMOD Int. Conf. Management of Data*, Seattle, Washington, 1998, pp. 94–105.

[17] C. C. Aggarwal, J. L. W. C. M. Procopiuc, P. S. Yu, and J. S. Park, "Fast algorithms for projected clustering," in *Proc. ACM-SIGMOD Intl. Conf. Management of Data*, 1999, pp. 61–72.

[18] C. Aggarwal and P. Yu, "Finding generalized projected clusters," in *Proc. ACM-SIGMOD Intl. Conf. Management of Data*, 2000, pp. 70–81.

[19] C. H. Cheng, A. W.-C. Fu, and Y. Zhang, "Entropy-based subspace clustering for mining numerical data," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM Press, 1999, pp. 84–93.

[20] C. Baumgartner, C. Plant, K. Kailing, H.-P. Kriegel, and P. Kröger, "Subspace selection for clustering high-dimensional data," in *Proc. 4th IEEE int. Conf. on Data Ming (ICDM 04)*, 2004, pp. 11–18.

[21] P. M. Murphy and D. W. Aha, "Uci repository of machine learning databases," 1994. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

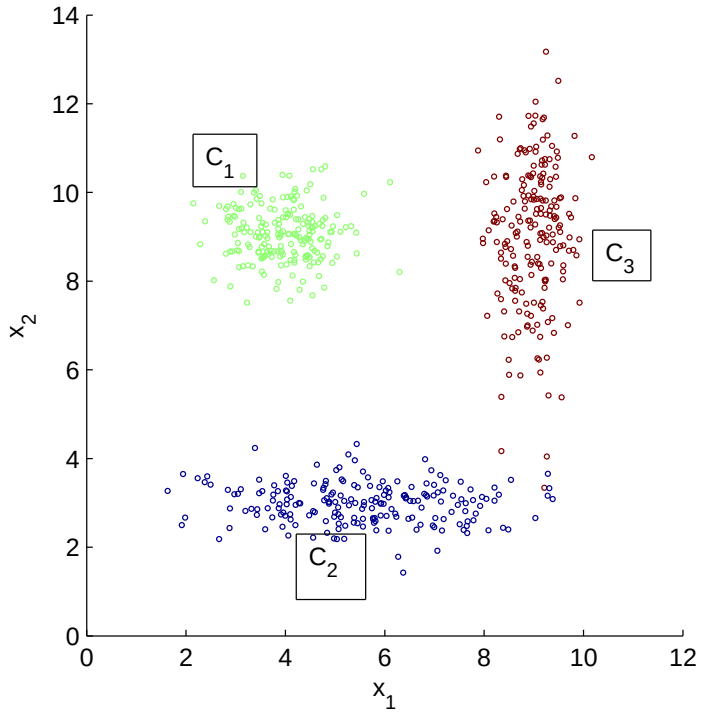


Figure 1: A three-cluster system with cluster C_1 embedded in feature set $\{x_1, x_2\}$, cluster C_2 embedded in feature subset $\{x_2\}$, and cluster C_3 embedded in feature subset $\{x_1\}$.

Table 1: Summary of the synthetic datasets, where N represents the number of patters, D the number of features, c the number of clusters, D_j the number of relevant feature respecting to the j-th cluster, and N_j the size of the j-th cluster.

Dataset	N	D	c	D_j	N_j
syn_1	600	15	3	3/3/3	200/200/200
syn_2	600	20	3	3/4/5	200/200/200
syn_3	1000	20	5	3/4/5/4/2	200/200/200/200/200
syn_4	900	30	3	3/3/3	200/300/400

Yuanhong Li, et. al.

Table 2: Results on the synthetic datasets. Saliency in the range [0, 1] is mapped to grey-scale [0, 255] linearly. For the clustering with localized feature saliency, each image is a mapping of feature saliency of one cluster, where rows and columns of pixels represent runs and features, respectively. The separated row pixels above an image represent the true relevant features. The global feature saliency is illustrated in the same way.

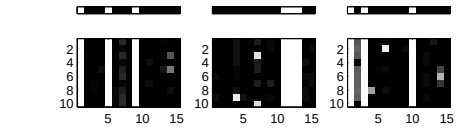
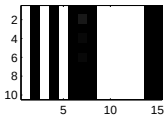
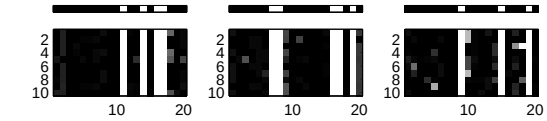
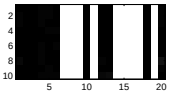
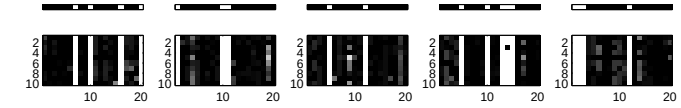
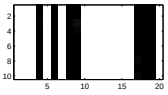
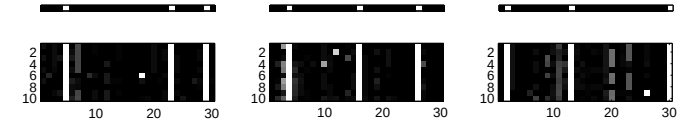
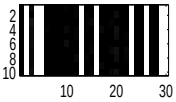
Dataset	Localized feature selection		Global feature selection	
	$\hat{c}$(std)	Saliency	$\hat{c}$(std)	Saliency
syn_1	3(0)		3(0)	
syn_2	3(0)		3(0)	
syn_3	5(0		5(0)	
syn_4			3(0)	

Table 3: Summary of UCI datasets

data	Description	N	D	c
wine	wine recognition	178	13	3
wdbc	Wisconsin diagnostic breast cancer	569	30	2
vehicle	vehicle classification	846	18	4
zernike	Zernike moments of digit images	2000	47	10

Table 4: Cluster numbers and pseudo error rates for UCI datasets.

data	Localized feature selection		Global feature selection	
	error (std)(%)	$\hat{c}$ (std)	error (std)(%)	$\hat{c}$ (std)
wine	2.1 (1.2)	3 (0)	2.4 (1.2)	3.3 (0.5)
wdbc	7.6 (0.6)	7.1 (0.7)	7.5 (1.2)	7.4 (0.8)
vehicle	44.6 (1.3)	9.2 (1.3)	45.4 (2.6)	10.5 (1.3)
zernike	44.9 (2.2)	15.3 (1.9)	47.6 (2.8)	16.7 (1.3)

Table 5: Feature saliency. Each image is a mapping of feature saliency for a cluster, with exception that the highlighted one represents the global feature saliency. Saliency values [0,1] are linearly mapped to grey-scale [0,255]. Each row represents a run, and each column represents a feature.

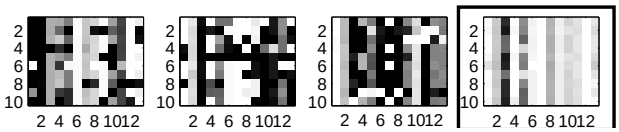
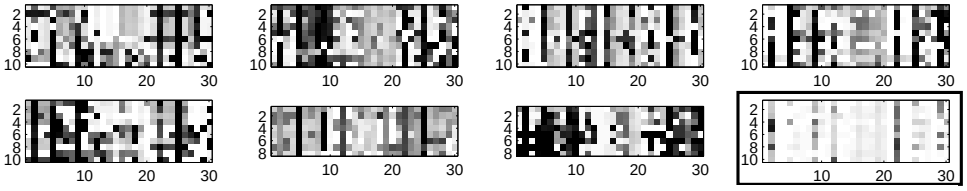
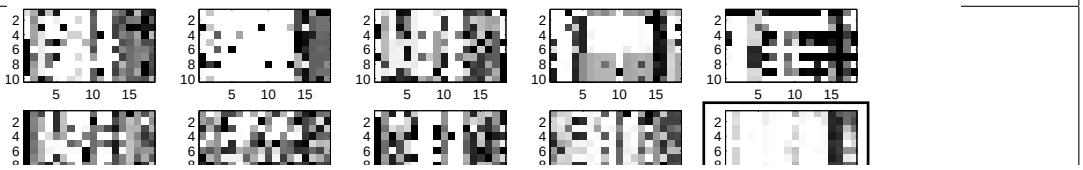
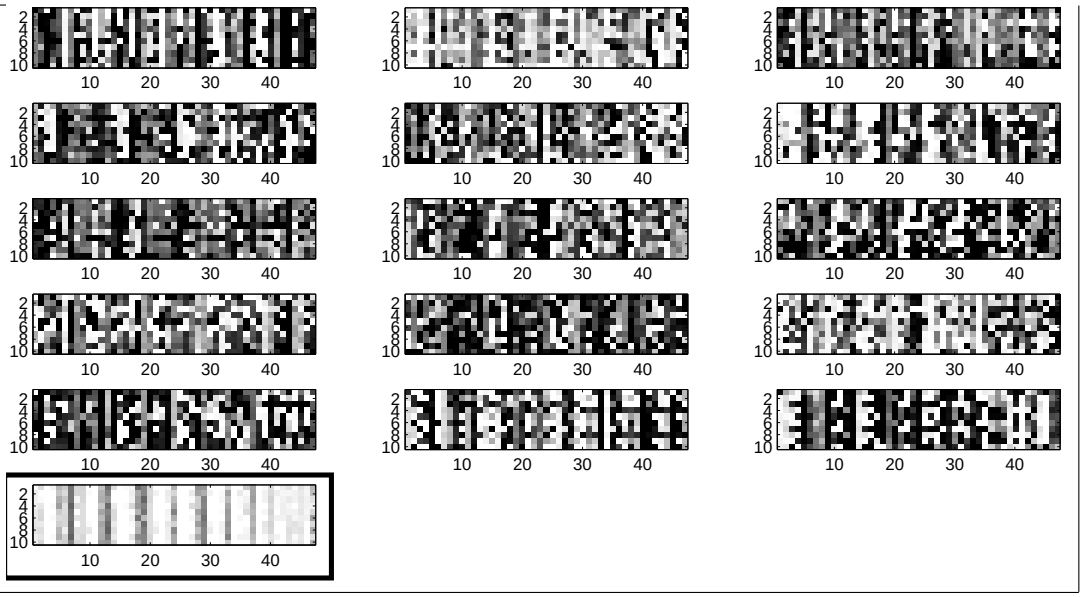
data	Feature saliency
wi	
wdbc	
vehicle	
zernike	

Table 6: Attributes for the Boston housing data.

Num.	Var.	Description
1	CRIM	per capita crime rate by town
2	ZN	land zoned for lots over 25,000 sq.ft.
3	INDS	proportion of non-retail business acres per town
4	CHAS	Charles River dummy variable
5	NOX	nitric oxides concentration
6	RM	number of rooms per dwelling
7	AGE	proportion of units built prior to 1940
8	DIS	distances to five Boston employment centres
9	RAD	accessibility to radial highways
10	TAX	full-value property-tax rate
11	PTRT	pupil-teacher ratio by town
12	B	$(Bk - 0.63)^2$ where Bk is the proportion of blacks
13	LSTT	% lower status of the population
14	MEDV	Median value of owner-occupied homes in $1000's

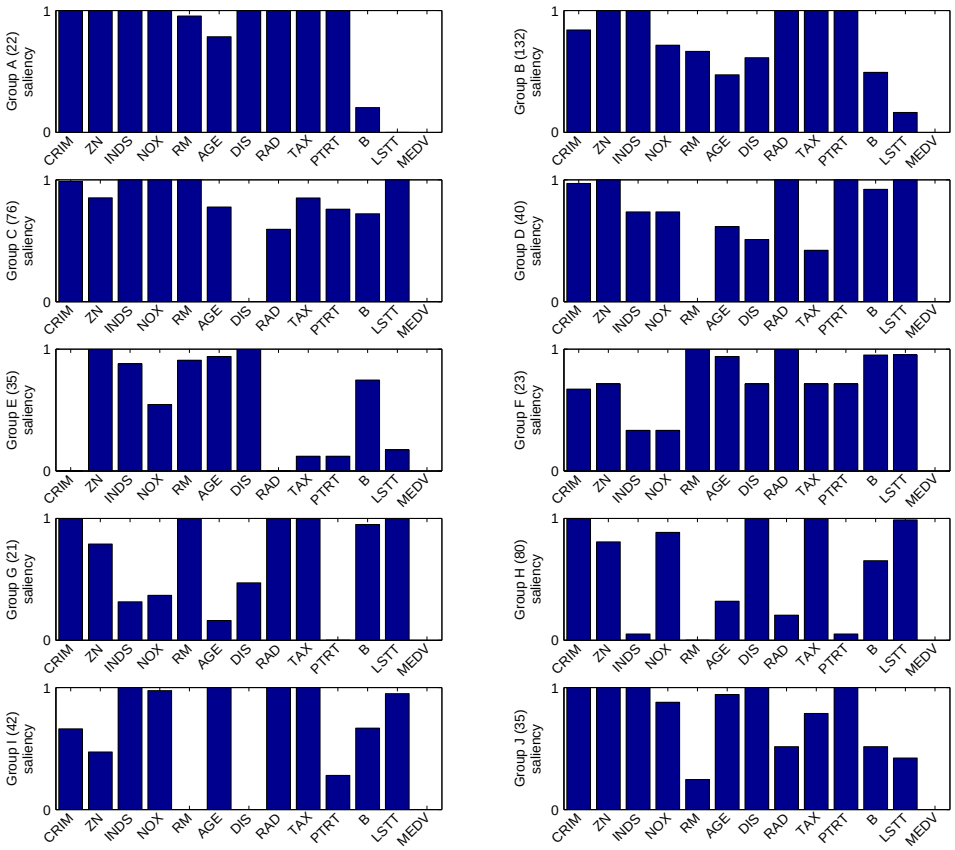


Figure 2: Localized feature saliency on the Boston housing dataset. The number of objects grouped together are listed with the group ID.

Yuanhong Li, et. al.