

Localized Feature Selection for Gaussian Mixtures Using Variational Learning

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

Yunqian Ma
Honeywell Labs, 1985 Douglas Drive North,
Golden Valley, MN 55422

Abstract

Typical unsupervised feature selection algorithms select a common feature subset for all the clusters. Consequently, clusters embedded in different feature subspaces are not discovered. In this paper, we propose a novel approach of simultaneous localized feature selection and model detection for unsupervised learning. In our approach, local feature saliency, together with other parameters of Gaussian mixtures, are estimated by Bayesian variational learning. Experiments performed on real-world datasets illustrate that our approach is superior over both global feature selection and subspace clustering methods.

1. Introduction

Clustering is the unsupervised classification of data objects into different groups (clusters) such that objects in one group are similar together and dissimilar from another group. A clustering algorithm typically considers all the available features to “learn” from data. In practice, however, some features can be irrelevant and therefore hinder the clustering performance, especially in a high-dimensional dataset. A viable solution is *feature selection*, a technique that chooses the “best” feature subset for clustering.

Feature selection has been extensively studied in supervised learning scenarios [6]. In unsupervised learning, feature selection becomes a more complex problem due to the unavailability of class labels. Algorithms in the literature are commonly categorized into two groups: filters and wrappers. A filter selects the relevant feature subset without the interaction with the mining algorithm, while a wrapper selects the relevant feature subset based on the performance of a particular mining algorithm. In the filter approach proposed in [3], an entropy measure is introduced, which is low if data has distinct clusters and high otherwise, and relevant

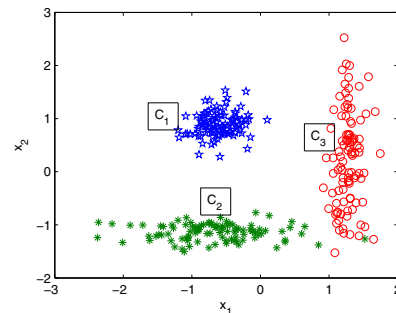


Figure 1. A three-cluster dataset with cluster C_1 embedded in feature set $\{x_1, x_2\}$, cluster C_2 embedded in feature subset $\{x_2\}$, and cluster C_3 embedded in feature subset $\{x_1\}$.

features are selected accordingly. The wrapper method presented in [4] evaluates the cluster quality over different feature subsets by normalizing cluster separability (for k -means clustering) or likelihood (for Expectation Maximization (EM) clustering) using the cross-projection method. The candidate feature subsets are generated by sequential forward search. In [7], feature saliency is estimated by the Maximum Likelihood (ML) or Maximum A Priori (MAP) with EM using Gaussian mixture models. Minimal Message Length (MML) is employed to estimate the number of components. [2] employs the same Gaussian mixture model as in [7] to describe feature relevance, but integrates model and feature selection under the Bayesian framework.

In general, unsupervised feature selection algorithms conduct feature selection in a *global* sense by producing a common feature subset for all the clusters. This, however, can be invalid in clustering practice, where the local intrinsic property of data matters more [8]. 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 can be grouped using $\{x_2\}$ and $\{x_1\}$, respec-

tively. A common feature subset, i.e., $\{x_1, x_2\}$, is unable to reflect the inherent structural properties of the three clusters. Apparently, clustering with *local* features is highly desired. Along this direction, bipartite graph partitioning [10] groups features together with patterns in every cluster. However, features are divided exclusively, which prevents the possibility of a feature being relevant to more than one cluster. Other approaches, usually referred as *subspace clustering* [9, 5], aim to seek low dimensional density areas embedded in a high dimensional feature space. Moreover, in real-world problems, the number of clusters is usually unknown, and hence needs to be detected. Note that different feature subsets may lead to different number of clusters. Feature selection and clustering model detection are strongly dependent [7]. This suggests that these two objectives must be achieved simultaneously. In this paper, we address the problem of simultaneous localized feature selection and model detection for unsupervised learning. We propose a novel localized Bayesian inference approach of Gaussian mixtures, which computes the local feature saliency, the number of clusters, and other parameters of a mixture through variational learning.

The rest of this paper is organized as follows: The Gaussian mixture model with local feature saliency is presented in Section 2. Variational learning is used in Section 3 to identify the mixture model and perform localized feature selection. We present our experimental results in Section 4, and conclusion in Section 5.

2. Mixture Model with Localized Feature Saliency

From a *model-based* perspective, each cluster can be mathematically represented by a parametric distribution. One of the most widely used distributions is the Gaussian. The entire dataset can therefore be modeled by a mixture of Gaussians. The clustering problem, thereby, reduces to a problem of estimating the parameters of the Gaussian mixture.

A finite mixture of densities with K components is represented by $p(y) = \sum_{j=1}^K \pi_j p(y|\theta_j)$, where π_j are called mixing coefficients, and θ_j are the parameters corresponding to component j . We use π to denote the set $\{\pi_j\}_{j=1, \dots, K}$, and similarly $\theta \equiv \{\theta_j\}_{j=1, \dots, K}$. We assume features are conditionally independent, and the importance of a feature can be different for different clusters. The feature relevance is represented by a matrix $S = \{s_{jl}\}_{K \times D}$, where $s_{jl} = 1$ indicates that feature l is associated with component j , otherwise, $s_{jl} = 0$. Let $\rho_{jl} = \Pr(s_{jl} = 1)$ be the probability that feature l is relevant to component j . Then, the like-

lihood can be obtained based on the following proposition¹.

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

$$p(y_i|\theta) = \sum_{j=1}^K \pi_j \prod_{l=1}^D (\rho_{jl} p(y_{il}|\theta_{jl}) + (1-\rho_{jl}) q(y_{il}|\lambda_{jl})).$$

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

The mixture component with localized feature saliency can be interpreted as follows: Assume that samples \mathcal{Y}_j are clustered to component j , with feature association indicator vector $(s_{jl})_{l=1, \dots, D}$. \mathcal{Y}_j are compact on feature subset \mathcal{F}_+ of which $s_{jl} = 1$, and are loosely distributed on feature subset \mathcal{F}_- of which $s_{jl} = 0$.

3. Estimate Localized Feature Selection by Variational Learning

The parameters of the above mixture model can be estimated by Maximum Likelihood (ML) with EM, or by Variational Learning of Bayesian approximation (VB). ML method treats the parameters as unknown but fixed, while VB places a prior probability on the parameters. These two algorithms usually produce identical results in many cases. However, in order to integrate cluster number estimation, ML method usually requires other criteria, such as Entropy measure and Minimal Message Length (MML) for optimization. For VB approach, this process can be implemented by the proper choice of prior probability over mixing coefficients. Another problem encountered by ML is that singular components lead to infinite likelihood, which is not the case with VB. We now present the VB approach to approximate the parameters in (1).

3.1 Variational Approximation

In general, to evaluate the likelihood of mixtures, conditioned on the mixing coefficients, we must marginalize the parameters as follows,

$$P(\mathcal{Y}|\pi) = \int P(\mathcal{Y}, \Theta|\pi) d\Theta \quad (1)$$

¹We skip the proof due to the space constraint.

where $\Theta \equiv \{\theta, z, S\}$ denotes all the parameters and latent variables. The integral sign denotes the joint integral over θ and the summation over z and S . This integral is analytically intractable. We therefore use variational methods to find the lower bound on $P(\mathcal{Y}|\pi)$. The idea of variational learning is to approximate $P(\Theta)$ with another easier distribution $Q(\Theta)$ by minimizing the Kullback-Leibler divergence $\text{KL}(Q||P)$ between Q and P . Assuming that $Q(\Theta)$ factorizes over subsets $\{\Theta_i\}$ of the variables in Θ , $Q(\Theta) = \prod_i Q_i(\Theta_i)$, the KL divergence can then be minimized over all possible factorial distributions by performing free-form minimization over Q_i ,

$$Q_i(\Theta_i) = \frac{\exp\langle \ln P(\mathcal{Y}, \Theta) \rangle_{k \neq i}}{\int \exp\langle \ln P(\mathcal{Y}, \Theta) \rangle_{k \neq i} d\Theta_i} \quad (2)$$

where $\langle \cdot \rangle_{k \neq i}$ denotes an expectation with respect to the distributions $Q_k(\Theta_k)$ for all $k \neq i$. (2) shows that the sufficient statistics of each distribution Q_i depends on the moments of other distributions $Q_{k \neq i}$, which implies an iterative solution for estimating the variational variables. In other words, with sufficient parameter initialization, the statistics can be updated by taking each factor in turn and replacing its sufficient statistics by the revised estimates.

3.2 Local feature saliency with variational learning

We now apply variational approach to Bayesian mixture of Gaussians with localized feature saliency. Given the sets of hidden variables $Z = \{z_{ij}\}$ and $S = \{s_{jl}\}$, the likelihood of the observed data is given by,

$$P(\mathcal{Y}|\pi, \mu, T, \rho, \epsilon, \gamma) = \prod_{i=1}^N \sum_{j=1}^K \pi_j \prod_{l=1}^D (\rho_{jl} \mathcal{N}(y_{il}|\mu_{jl}, \tau_{jl}) + (1 - \rho_{jl}) \mathcal{N}(y_{il}|\epsilon_{jl}, \gamma_{jl})). \quad (3)$$

where $\mu = \{\mu_{jl}\}$ and $T = \{\tau_{jl}\}$ denote the means and inverse variances of the ‘‘useful’’ subcomponents, while $\epsilon = \{\epsilon_{jl}\}$ and $\gamma = \{\gamma_{jl}\}$ are the set of parameters for the ‘‘noisy’’ subcomponents. The distribution of the hidden variable Z given the mixing probabilities $\pi = \{\pi_j\}$ and the distribution of the hidden variable S given the mixing probabilities $\rho = \{\rho_{jl}\}$ are governed as, $P(Z|\pi) = \prod_{i=1}^N \prod_{j=1}^K \pi_j^{z_{ij}}$, and $P(S|\rho) = \prod_{i=1}^N \prod_{j=1}^K \prod_{l=1}^D \rho_{jl}^{s_{jl}^{(i)}} (1 - \rho_{jl})^{1-s_{jl}^{(i)}}$. Model selection is accomplished by introducing conjugate priors over the means and inverse covariances, $P(\mu) = \prod_{j=1}^K \prod_{l=1}^D \mathcal{N}(\mu_{jl}|m_l, c)$ and $P(T) = \prod_{j=1}^K \prod_{l=1}^D \Gamma(\tau_{jl}|\alpha, \beta)$, where m_l, c, α, β are hyperparameters which control the prior distributions, and $\Gamma(\cdot)$ is the gamma distribution.

Applying (2) to the above Bayesian model, we get,

$$Q_Z(Z) = \prod_{i=1}^N \prod_{j=1}^K r_{ij}^{z_{ij}} \quad (4)$$

$$Q_\mu(\mu) = \prod_{j=1}^K \prod_{l=1}^D \mathcal{N}(\mu_{jl}|m_{jl}^v, c_{jl}^v) \quad (5)$$

$$Q_T(T) = \prod_{j=1}^K \prod_{l=1}^D \Gamma(\tau_{jl}|\alpha_{jl}^v, \beta_{jl}^v) \quad (6)$$

$$Q_S(S) = \prod_{i=1}^N \prod_{j=1}^K \prod_{l=1}^D \omega_{ijl}^{s_{jl}^{(i)}} (1 - \omega_{ijl})^{1-s_{jl}^{(i)}} \quad (7)$$

where $r_{ij}, m_{jl}^v, c_{jl}^v, \alpha_{jl}^v, \beta_{jl}^v$, and ω_{ijl} are variational parameters that determine the density involved in Q . We take an EM-like process to approximate them by minimizing $\text{KL}(Q||P)$ w.r.t Q at the first step, and minimizing $\text{KL}(Q||P)$ w.r.t. parameters $\pi_j, \rho_{jl}, \epsilon_{jl}$, and γ_{jl} at the second step.

This model has a property that the components with similar parameters fitting the same Gaussian will compete each other, yielding a dominant cluster. Thus, we can initialize the model with a large number of clusters, and eliminate the trivial clusters during the iterations. Finally, the algorithm will produce a model with localized feature saliency represented by ρ_{jl} , and identify the number of clusters simultaneously.

3.3 Advantages of the proposed approach

The advantages of our approach are summarized as follows:

1. Compared with global methods, our method can reveal cluster-wise feature relevance, thus provides users more accurate information about the underlying model which generates the data.
2. Compared with subspace clustering methods, our method does not require users to provide parameters which are critical but in practice almost impossible to be set in advance, such as the number of clusters, the density threshold, and the desired dimensionality.
3. Our method avoids heuristical navigation over the large pool of possible feature subsets. The computational cost for each iteration of the proposed algorithm is $\mathcal{O}(KND)$. It does not grow exponentially with D or N , thus can be scaled to large datasets. The total computational time depends on the number of iterations required for converging.

4. Experimental Results

To thoroughly evaluate the proposed Localized Feature Selection using Variational Bayesian (LFSVB) al-

Table 1. Mutual information I and the estimated cluster number \hat{c} , represented by mean and standard deviation over 10 different runs, on UCI datasets. For COSA, the number of clusters is determined manually, indicated by “*”.

Data	Algo	\hat{c} (std)	I (std)
Heart	LFSVB	2.8(0.8)	0.15(0.07)
	COSA	2*	0.21 (0.01)
	GFSVB	3(0.7)	0.09(0.06)
Ion	LFSVB	3.8(1.1)	0.33 (0.1)
	COSA	4*	0.30(0.01)
	GFSVB	3.4(0.9)	0.21(0.05)
Vehicle	LFSVB	9.9(1.7)	0.63 (0.05)
	COSA	9*	0.48(0.01)
	GFSVB	10.5(1.5)	0.58(0.09)
Wine	LFSVB	3.1(0.3)	1.44 (0.07)
	COSA	3*	1.26(0.01)
	GFSVB	3.4(0.7)	1.42(0.06)
WDBC	LFSVB	6.3(0.8)	0.68 (0.02)
	COSA	10*	0.59(0.01)
	GFSVB	7.6 (0.9)	0.67(0.02)
Yeast	LFSVB	11.4(2.1)	0.39 (0.06)
	COSA	13*	0.15(0.02)
	GFSVB	6.8(0.8)	0.36(0.01)

gorithm, we have compared it with the global feature selection approach in [2] (GFSVB) (also based on variational learning), and subspace clustering methods in [5] (COSA, which produces soft feature importance), on six real-world datasets downloaded from UCI Machine Learning Repository [1], namely, Heart Disease, Ionosphere, Vehicle, Wine, WDBC, and Yeast. The mutual information (I) between the true labels and cluster labels is used to evaluate the performance of different algorithms. A higher value of I indicates that the cluster results are closer to the true class group.

Table 1 shows the mean and standard deviation of the cluster numbers and mutual information over 10 runs for the three algorithms. Note that cluster numbers for COSA are set manually based on the dendrogram. On the average mutual information, LFSVB outperforms GFSVB on five (out of six) datasets (Heart, Ion, Vehicle, Wine, and Yeast). On WDBC, it is as good as GFSVB. LFSVB also outperforms COSA on five (out of six) datasets (Ion, Vehicle, Wine, WDBC and Yeast).

In addition, different relevant feature subsets are selected by LFSVB for different clusters, whose sizes are usually smaller than the global relevant feature subset. For example, when the cut-off threshold is set to 0.5, the feature subsets associated with three clusters in the Heart dataset are $C_1 : [1, 4, 5, 8, 10]$, $C_2 : [1]$ and $C_3 : [4, 5, 8]$, while the

global one is $[4, 5, 8, 12]$. For the Wine dataset, the three subsets are $C_1 : [3, 4, 6, 7, 8, 12, 13]$, $C_2 : [2]$ and $C_3 : [1, 2, 4, 10, 12, 13]$ and the global one is $[1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13]$. These results indicate that a globally relevant feature can be irrelevant to some clusters. Thus, clustering with localized feature selection algorithm provides users with more accurate knowledge regarding the underlying model from which the cluster component is generated.

5. Conclusion

In this paper, we propose a fully-automated method to identify useful patterns embedded in feature subspaces by integrating local feature selection, model detection, and clustering into a unified Bayesian framework through variational learning. We demonstrate the advantages of our algorithm over global feature selection and subspace clustering methods on several real-world datasets.

Acknowledgment

This research was partially funded by National Science Foundation under grants: IIS-0713315 and CNS-0751045, and by the 21st Century Jobs Fund Award, State of Michigan, under grant: 06-1-P1-0193.

References

- [1] A. Asuncion and D. Newman. UCI repository of machine learning databases, 2007.
- [2] C. Constantinopoulos, M. K. Titsias, and A. Likas. Bayesian feature and model selection for gaussian mixture models. *PAMI*, 28(6):1013–1018, 2006.
- [3] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering - a filter solution. In *ICDM*, pages 115–122, 2002.
- [4] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *J. Machine Learning Research*, 5:845–889, 2004.
- [5] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *J. Royal Statistical Society: Series B*, 66(4):815–849, 2004.
- [6] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *PAMI*, 19(2):153–158, 1997.
- [7] M. H. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *PAMI*, 26(9):1154–1166, 2004.
- [8] Y. Li, M. Dong and J. Hua. Localized Feature Selection for Clustering. *Pattern Recognition Letters*, 29:10–18, 2008.
- [9] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl.*, 6(1):90–105, 2004.
- [10] H. Zha, X. He, C. Ding, M. Gu, and H. Simon. Bipartite graph partitioning and data clustering. In *ACM CIKM*, pages 25–32, 11 2001.