Simultaneous Non-gaussian Data Clustering, Feature Selection and Outliers Rejection

Nizar Bouguila¹, Djemel Ziou², and Sabri Boutemedjet²

¹ Concordia Institute for Information Systems Engineering, Concordia University,
Montreal, Canada, Qc, H3G 2W1

² Département d'Informatique, Université de Sherbrooke, Canada, Qc, J1K2R1
bouguila@ciise.concordia.ca,
{djemel.ziou,sabri.boutemedjet}@usherbrooke.ca

Abstract. A method for simultaneous non-Gaussian data clustering, feature selection and outliers rejection is proposed in this paper. The proposed approach is based on finite generalized Dirichlet mixture models learned within a framework including expectation-maximization updates for model parameters estimation and minimum message length criterion for model selection. Through a challenging application involving texture images discrimination, it is demonstrated that the developed procedure performs effectively in avoiding outliers and selecting relevant features.

1 Introduction

Clustering is an important problem in data mining, pattern recognition, computer vision and machine learning applications. It is well-known that this problem becomes more challenging in the presence of outliers and irrelevant features which may compromise the final clustering models [1]. A good clustering model should be sensitive to the extracted features but not to the noise (i.e outliers). A challenging problem in this case is to determine if all the features are necessary and relevant for the clustering task. Over the last few years, there has been a surge of interest in unsupervised feature selection. While the volume of literature on the subject is notably smaller than the literature related to the supervised case, there is nonetheless a steady stream of interesting approaches (see, for instance, [2,3]). All these approaches, however, assume noiseless data that contain only inliers. The approach proposed in [2] was among the earliest, but was based on two main assumptions namely the independency of the features supposed to follow Gaussian distributions. These assumptions have been relaxed then in one of our previous works using generalized Dirichlet mixture models [3].

The current paper sets out to add robustness to the unsupervised feature selection approach that we have proposed in [3] by automatically detecting potential outliers. This is well-known to be a challenging condition for all learning algorithms which may have significant impact on the effectiveness of statistical modeling. The main goal is to accurately identify outliers and remove their effect from the rest of the model fitting. It is worth commenting that the main contribution of the proposed work is that the outliers effects are estimated simultaneously with the features weights and the model's parameters.

The rest of this paper is organized as follows. After introducing our model and justifying its use and how it can be learned in Section 2, we illustrate the proposed approach on a real application which concerns texture images discrimination in Section 3. Finally, we conclude the paper with a summary of the work in Section 4.

2 Unsupervised Feature Selection and Outliers Rejection

2.1 The Model

Let us consider a data set of N vectors $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$, where each $\mathbf{X}_i = (X_{i1}, \dots, X_{iD})$ is a D -dimensional vector of features representing a given image, for instance. Set of vectors generally contains examples that belong to many clusters and can be modeled by a finite mixture of distributions

$$p(\mathbf{X}_i|\Theta) = \sum_{j=1}^M p_j p(\mathbf{X}_i|\theta_j) \quad (1)$$

where $p_j > 0$, are the mixing proportions, M is the number of mixture components, and $\Theta = \{\mathbf{P} = (p_1, \dots, p_M), \boldsymbol{\theta} = (\theta_1, \dots, \theta_M)\}$ is the set of parameters in the mixture model. A critical problem in this case is the choice of the probability density function $p(\mathbf{X}_i|\theta_j)$ to represent each component. In [3], we have shown that the generalized Dirichlet distribution offers the unusual luxury that by a simple transformation we go from one space to another one where all the variables are independent and follow Beta distributions [4,5]. This allows us to easily formalize the fact that different features may have different weights by approximating a generalized Dirichlet mixture as following [3]

$$p(\mathbf{X}_i|\Theta) = \sum_{j=1}^M p_j \prod_{d=1}^D (\rho_d p(X_{id}|\alpha_{jd}, \beta_{jd}) + (1 - \rho_d) p(X_{id}|\alpha_d, \beta_d)) \quad (2)$$

where $\Theta = \{\mathbf{P}, \{\rho_d\}, \{\alpha_{jd}, \beta_{jd}\}, \{\alpha_d, \beta_d\}\}$ is the set of all the model parameters, ρ_d represents the probability that the d^{th} feature is relevant for clustering, $p(X_{id}|\alpha_{jd}, \beta_{jd})$ is a Beta distribution that represents a feature X_{id} when it is relevant, and $p(X_{id}|\alpha_d, \beta_d)$ is an univariate Beta with parameters α_d, β_d and can be viewed as a common background distribution to explain nonsalient features. It is noteworthy that in the previous model no knowledge is available as to which vector $\mathbf{X}_i$ is not representative (i.e. an outlier) and then is not really generated from our assumed statistical model. Model selection, feature selection and parameters estimation are generally dramatically affected by outliers. Thus, it is crucial to extend the model in Eq. 2 to handle outliers. We approach this problem by incorporating an auxiliary outlier component to which we associate a uniform density¹ into the model:

¹ Supposing that the outliers follow a uniform distribution is a reasonable common assumption that have been considered in several computer vision and pattern recognition applications (see, for instance, [6,7]).

$$p(\mathbf{X}_i|\Theta) = \sum_{j=1}^M p_j \prod_{d=1}^D (\rho_d p(X_{id}|\alpha_{jd}, \beta_{jd}) + (1 - \rho_d)p(X_{id}|\alpha_d, \beta_d)) + p_{M+1}U(\mathbf{X}_i) \quad (3)$$

where $p_{M+1} = 1 - \sum_{j=1}^M p_j$ is the probability that $\mathbf{X}_i$ was not generated by the central mixture model and $U(\mathbf{X}_i)$ is a uniform distribution common for all data to model isolated vectors which are not in any of the M clusters and which show significantly less differentiation among clusters. The previous model can be viewed as a way to robustify unsupervised feature selection to learn the right meaning from the right observations (i.e inliers). Notice that when $p_{M+1} = 0$ the outlier component is removed and the previous equation is reduced to Eq. 2.

2.2 Model Learning

A well-known approach for unknown parameters estimation is the technique of maximum likelihood (ML) which properties have been extensively examined in the past. Using ML the parameters are estimated by maximizing the log-likelihood function as following

$$\hat{\Theta} = \arg \max_{\Theta} \left\{ \log p(\mathcal{X}|\Theta) = \sum_{i=1}^N \log \left[\sum_{j=1}^M \left(p_j \prod_{d=1}^D (\rho_d p_{jd}(X_{id}) + (1 - \rho_d)p(X_{id})) \right) + p_{M+1}U(\mathbf{X}_i) \right] \right\} \quad (4)$$

where $p_{jd}(X_{id}) = p(X_{id}|\alpha_{jd}, \beta_{jd})$ and $p(X_{id}) = p(X_{id}|\alpha_d, \beta_d)$, which gives us

$$p_j = \frac{\sum_{i=1}^N p(j|\mathbf{X}_i)}{N} \quad j = 1, \dots, M+1 \quad (5)$$

where

$$p(j|\mathbf{X}_i) = \begin{cases} \frac{p_j \prod_{d=1}^D (\rho_d p_{jd}(X_{id}) + (1 - \rho_d)p(X_{id}))}{\sum_{j=1}^M \left(p_j \prod_{d=1}^D (\rho_d p_{jd}(X_{id}) + (1 - \rho_d)p(X_{id})) \right) + p_{M+1}U(\mathbf{X}_i)} & \text{if } j = 1, \dots, M \\ \frac{p_{M+1}U(\mathbf{X}_i)}{\sum_{j=1}^M \left(p_j \prod_{d=1}^D (\rho_d p_{jd}(X_{id}) + (1 - \rho_d)p(X_{id})) \right) + p_{M+1}U(\mathbf{X}_i)} & \text{if } j = M+1 \end{cases} \quad (6)$$

is the posterior probability that a vector $\mathbf{X}_i$ will be considered as an inlier and then assigned to a cluster $j, j = 1, \dots, M$ or as an outlier and then affected to cluster $M+1$ which allows to safeguard against erroneous feature selection. Details about the estimation of the other parameters namely $\alpha_{jd}, \beta_{jd}, \alpha_d, \beta_d$ and ρ_d can be found in [3]. It is desirable to determine the simplest model that can explain the data accurately. It is noteworthy that the simplicity is measured in our case by the number of mixture components and the number of relevant features. The simplest model can be viewed as the one that maximizes the message length, to reach an acceptable balance between model complexity and goodness

of fit. Details and discussions about the computation of the message length can be found in [4,3].

Algorithm

For each candidate value of M :

1. Set $\rho_d \leftarrow 0.5$, $d = 1, \dots, D$, $j = 1, \dots, M$ and initialization of the rest of parameters².
2. Iterate the two following steps until convergence:
 - (a) E-Step: Update $p(j|\mathbf{X}_n)$ using Eq. 6.
 - (b) M-Step: Update the p_j , ρ_d , β_{jd} , α_{jd} , β_d and α_d as proposed in [3].
3. Calculate the associated message length [3].
4. Select the optimal model that yields the smallest message length.

3 Experimental Results: Texture Discrimination

The problem of texture discrimination has been the subject of several studies in the past (see, for instance, [8,9]). Several texture descriptors have been proposed [10] and applied to several problems such as industrial surface inspection and content-based image retrieval. We shall not elaborate further on the different approaches that have been proposed to model textures which is clearly beyond the scope of this paper. Rather we focus on an important texture feature extraction approach that has received a lot of attention recently. This approach is based on the representation of texture using visual dictionary obtained through the quantization of the appearance of local regions, described by local features, which gives characteristic texture elements generally called *textons* (see, for instance, [11]).

For our experimental evaluation, we use the KTH-TIPS [12] texture data set which is composed of 10 texture classes with 81 images per class (see Fig. 1). We add to this data set 20 different images, taken from different classes of the MIT vistex texture data set, which will be considered as outliers (see Fig. 2). An important step in our application is the extraction of local features to describe the textures. In our case we use the approach proposed in [11] which uses $n \times n$ pixel compact neighborhoods as image descriptors. Using this approach each texture pixel is described by an n^2 -dimensional vector which represents the pixel intensities of its $n \times n$ square neighborhood. Following [11], we use $n = 7$, thus

Fig. 1. Examples of images from the 10 different classes in the KTH-TIPS data set

² The initialization is based on the K-Means algorithm and the method of moments by considering that $M + 1$ clusters are present in the data.

Fig. 2. Sample images from the Vistex data set

each pixel is represented by a 49-dimensional vector. Then, we construct a global *texon* vocabulary via the clustering of the different descriptors. In particular, we extract 10 textons using K-Means (i.e. the textons are actually the K-Means cluster centers) for each texture class and then concatenate the textons of the different classes to form the visual vocabulary. Thus, the vocabulary size is 100. Once the vocabulary of textons is built, each texture image can be represented then by a vector of frequencies which we normalize.

Evaluation results by considering different clustering scenarios and by comparing generalized Dirichlet mixture (GDM) and Gaussian mixture (GM) are summarized in table 1. It is noteworthy that in both cases (GDM or GM), we were able to find the exact number of clusters only when we have rejected the outliers. According to the results in table 1 it is clear that the GDM outperforms the GM which is actually an expected result taking into account the fact that the features vectors are proportional vectors for which the GDM is one of the best choices. It is clear also that feature selection improves the clustering performance especially when combined with outliers rejection.

Table 1. Clustering results for the KTH-TIPS data set by considering different scenarios. Clust denotes clustering without feature selection and outliers rejection, FS denotes feature selection and OR denotes outliers rejection.

	Clust	Clust+FS	Clust+OR	Clust+FS+OR
GDM	72.23%	74.88%	74.31%	76.77%
GM	68.76%	70.13%	70.02%	72.22%

4 Conclusion

The goal of the proposed work is to tackle simultaneously two of the chief obstacles to effective clustering namely the presence of outliers and irrelevant features. Our approach is based on extending the unsupervised feature selection technique that we have previously proposed in [3] to take into account outliers. The merits of the proposed framework have been shown via the clustering of texture images.

Acknowledgment

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proc. of the Annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 82–88 (1996)
2. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(9), 1154–1166 (2004)
3. Boutemedjet, S., Bouguila, N., Ziou, D.: A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(8), 1429–1443 (2009)
4. Bouguila, N., Ziou, D.: High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10), 1716–1731 (2007)
5. Bouguila, N., Ziou, D.: A New Approach for High-Dimensional Unsupervised Learning: Applications to Image Restoration. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) *PREMI 2005*. LNCS, vol. 3776, pp. 200–205. Springer, Heidelberg (2005)
6. Williams, C.K.I., Titsias, M.K.: Greedy Learning of Multiple Objects in Images using Robust Statistics and Factorial Learning. *Neural Computation* 16(5), 1039–1062 (2003)
7. Sudderth, E.B., Torralba, A., Freeman, W.T., Wilsky, A.S.: Depth from Familiar Objects: A Hierarchical Model for 3D Scenes. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2410–2417 (2006)
8. Luettgen, M.R., Willsky, A.S.: Likelihood Calculation for a Class of Multiscale Stochastic Models, with Application to Texture Discrimination. *IEEE Transactions on Image Processing* 4(2), 194–207 (1995)
9. Ojala, T., Valkealahti, K., Oja, E., Pietikäinen, M.: Texture Discrimination with Multidimensional Distributions of Signed Gray-Level Differences. *Pattern Recognition* 34, 727–739 (2001)
10. Ojala, T., Mäenpää, T., Viertola, J., Kyllönen, J., Pietikäinen, M.: Empirical Evaluation of MPEG-7 Texture Descriptors with a Large-Scale Experiment. In: Proc. of the 2nd International Workshop on Texture Analysis and Synthesis, pp. 99–102 (2002)
11. Varma, A., Zisserman, A.: Texture classification: are filter banks necessary? In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 691–698 (2003)
12. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.-O.: On the significance of real-world conditions for material classification. In: Pajdla, T., Matas, J.(G.) (eds.) *ECCV 2004*. LNCS, vol. 3024, pp. 253–266. Springer, Heidelberg (2004)