A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection

Nizar Bouguila a,⁎, Khaled Almakadmeh a, Sabri Boutemedjet b

a Concordia Institute for Information Systems Engineering, Faculty of Engineering and Computer Science, Concordia University, Montreal, Quebec, Canada H3G 2W1
b Département Informatique, Université de Sherbrooke, Sherbrooke, Quebec, Canada

A R T I C L E   I N F O

Keywords:
Clustering
Finite mixture models
Maximum likelihood
EM
Feature selection
Outlier rejection
Gamma distribution
Integrated likelihood
Shape modeling

A B S T R A C T

Model-based approaches and in particular finite mixture models are widely used for data clustering which is a crucial step in several applications of practical importance. Indeed, many pattern recognition, computer vision and image processing applications can be approached as feature space clustering problems. For complex high-dimensional data, however, the use of these approaches presents several challenges such as the presence of many irrelevant features which may affect the speed and also compromise the accuracy of the used learning algorithm. Another problem is the presence of outliers which potentially influence the resulting model’s parameters. For this purpose, we propose and discuss an algorithm that partitions a given data set without a priori information about the number of clusters, the saliency of the features or the number of outliers. We illustrate the performance of our approach using different applications involving synthetic data, real data and objects shape clustering.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

A recurring subject in pattern recognition and computer vision applications is the separation of data (images, videos, objects, etc.) into homogeneous clusters (Jain, Murty, & Flynn, 1999). This topic has been extensively studied and different approaches and algorithms have been proposed and applied to several problems. Generally, an important step in these problems is the representation of a given image by a vector of features which is generally high-dimensional (Trunk, 1979). Although different, the majority of approaches agree that a good clustering model should be sensitive to the extracted features but not to the noise (i.e. outliers) which may be present. A challenging problem in this case is to determine if all the features are necessary and relevant for the clustering task (Fukunaga, 1990; Hu, Pedrycz, Yu, & Lang, 2010). Many methods have been developed to estimate the usefulness of features for clustering and prediction. Reducing the number of features not only speeds up the learning and training process, but also prevents over-fitting allowing the generation of the most optimal model to represent the data and to reflect its regularities. Feature selection is sensitive to the choice of the number of clusters describing the data. Indeed, if the selected number of clusters is very incorrect, the feature selection may be inaccurate as well. Moreover, it is often the case that some of the data is not representative and may deteriorate clustering performance. Thus, it is crucial to automatically detect these data commonly called outliers and which can be described1 as the observations that do not come from the model (Barnett & Lewis, 1994).

To achieve maximum utility and flexibility, finite mixture models are widely used and are well-known for their efficiency in clustering heterogenous data sets. In mixture model, data are supposed to be described by a number of distributions mixed in varying proportions. A fundamental element when using finite mixture models is then the choice of the components densities which has to take into account the characteristics of the data. Works dealing with finite mixtures vary in the assumptions that they make about the mixture distributions. Generally, data are assumed to be normally distributed. The normal distribution is, however, not always appropriate in pattern recognition, signal and image processing applications.2 One of the drawbacks is the rigidity of its shape which prevents it from yielding a good modeling and adequate representation of actual data (Oja, 1981). Thus, it is important to take into account the underlying structural characteristics of the data domain and the nature of the patterns that we would like to discover.

⁎ Corresponding author.
E-mail addresses: bouguila@ciise.concordia.ca (N. Bouguila), k_almakh@encs.concordia.ca (K. Almakadmeh), sabri.boutemedjet@usherbrooke.ca (S. Boutemedjet).

1 There are many other definitions that have been discussed, for instance, in D. Muñoz-Garcia, Moreno-Rebollo, and Pascual-Acosta (1990).
2 For instance, previous studies of natural images have revealed that their statistics are not Gaussian. Such studies suggest the use of more flexible models (see, for example, Srivastava, Lee, Simoncelli, & Zhu (2003)).
This paper addresses unsupervised learning of data defined over the positive reals \([0, \infty]\), which naturally appear in several image processing, pattern recognition and computer vision applications, and for which the Gamma distribution is known to be a good flexible choice and an accurate alternative to the Gaussian (Hsiao, Rangarajan, & Gindi, 2002; Petroiu, Giorgini, & Smits, 2002; Petroiu & Matrouccoli, 1998; Samadani, 1995; Zaart, Ziou, Wang, & Jiang, 2002; Ziou & Bouguila, 2004; Ziou, Bouguila, Allili, & El Zaart, 2009). We propose then a novel statistical mixture model, based on the Gamma distribution,\(^3\) which simultaneously detects outliers, by explicitly introducing a new class to model outliers, and select features by explicitly introducing a common background distribution to explain nonsalient features. Both the outlier component, the background distribution and their respective weights are repeatedly adjusted so as to maximize the total integrated likelihood of the model.

Determining the relevant features is one of central problems in machine learning and pattern recognition and several approaches have been proposed in the past. Most of the approaches, however, have focused on the supervised case which is justified by the difficulty to assess feature relevancy without resorting class labels and more challenge is added when the number of clusters is unknown. The literature about supervised feature selection is considerable (see, for instance, Wettschereck, Aha, & Mohri (1997)). Recently the literature about supervised feature selection is considerable (see, for instance, Wettschereck, Aha, & Mohri (1997)). Recently, the authors in Law, Figueiredo, and Jain (2004) proposed a simultaneous feature selection and clustering algorithm using finite Gaussian mixture models, that assume independence of features, and minimum message length (MML) criterion. The MML criterion was used also in Boutemedjet, Bouguila, and Ziou (2009) where the authors proposed an unsupervised approach for feature selection and extraction to model non-Gaussian data. The author in Bouguila (2009) proposed a probabilistic approach that assigns relevancy weights to discrete features that are considered as random variables modeled by finite discrete mixture models learned using stochastic complexity. In Graham and Miller (2006) an algorithm for subspace clustering based on expectation–maximization and using a minimum description length (MDL) criterion was proposed to select the relevant features subset and the number of clusters which best describes a given image. Set of vectors generally contains examples that belong to many clusters and can be modeled by a finite mixture of distributions

\[
p(X_i|\Theta_M) = \sum_{j=1}^{M} p_j p_j(X_i|\Theta_j),
\]

where \(p_j > 0\), are the mixing proportions, \(M\) is the number of mixture components, and \(\Theta_M = \{P = (p_1, \ldots, p_M), \theta = (\theta_1, \ldots, \theta_M)\}\) is the set of parameters in the mixture model. Eq. (1) constitutes actually a family of models which can be viewed as additive models in statistics (Hastie & Tibshirani, 1990). A critical problem in this case is the choice of the probability density function to represent each component. In several applications the vectors elements \(X_{d,j}\), \(d = 1, \ldots, D\) are positive. In this case one of the most useful probability density functions is the multivariate Gamma. A simple multivariate widely used form, assuming independence of variables, is given as the following (Webb, 2000, 2003)\(^5\):

\[
p(X_i|\theta_j) = \prod_{d=1}^{D} p(X_{d,j}|\theta_{d,j}) = \prod_{d=1}^{D} X_{d,j}^{\beta_d-1} \exp\left(-\frac{X_{d,j}}{\beta_d}\right),
\]

where \(\theta_j = (\beta_1, \ldots, \beta_D)\) is the univariate Gamma distribution, \(\beta_d > 0\), \(\beta_d > 0\), \(d = 1, \ldots, D\), and represent shape and scale parameters, respectively, and \(\Gamma(.)\) denotes the Gamma function.

As we have mentioned in the introduction, an important step in pattern recognition applications in general and clustering in particular is feature selection. The main objective is to choose only those features which are better suited and relevant for the problem under study (i.e. features should be selected according to their discrimination power). As a classic problem, feature selection has been defined in different ways in the literature (see, for instance, Das (1971) and Liu & Motoda, 2008). A widely used approach to perform feature selection, in order to take into account the fact that different features may have different weights, in the case of finite mixture models is to

\(^3\) This distribution has a long history. For instance, in 1895, Karl Pearson directed attention to the Gamma as a model for skewed data (Pearson, 1895).

\(^4\) For instance, criteria for the rejection of outliers have been proposed when dealing with univariate Gaussian observations in Peirce (1852), Irwin (1925), and Pearson and Sekar (1936). An approach in the case of univariate Gamma samples has been proposed in Lewis and Fieller (1979), also.

\(^5\) It is noteworthy that some techniques have been proposed in the case of subspace clustering and feature extraction (see, for instance, Ke & Kanade (2004) and Hubert & Engelen (2004)), which are beyond the intended scope of this paper.

\(^6\) Many parameterizations do exist for the Gamma distribution (Webb, 2002). In this paper we consider the parametrization given in Eq. (2).
approximate Eq. (1) as following (Boutemedjet et al., 2009; Graham & Miller, 2006; Law et al., 2004; Novovičová, Pudil, & Kittler, 1996)

\[
p(X_i|\Theta) = \frac{\rho(X_i|\Theta)}{\rho_0} + (1 - \rho_0)p(X_i|\lambda_{jd}),
\]

where \( \Theta = \{\Theta_{0k}, \rho_{kd}, \lambda_{jd}\} \) is the set of all the model parameters, \( \rho_{kd} \) represents the probability that the \( d \)th feature is relevant for clustering, and \( p(X_i|\lambda_{jd}) \) is an univariate Gamma with parameters \( \lambda_{jd} = (\lambda_{jkd}, \beta_{jkd}) \) and can be viewed as a common background distribution to explain nonsalient features. Notice that if \( \rho_{kd} = 0 \), \( d = 1, \ldots, D \), the model in Eq. (3) is reduced to the one in Eq. (1). Notice also that feature saliency is defined globally (i.e. a given feature is relevant or not to all the mixture components) which can be an invalid assumption in practical clustering problems as shown in Li et al. (2009). A better approach to take into account the local intrinsic property of the data, which plays an important role (Li & Hua, 2008), by assuming that the relevancy of features is different for different classes which can be modeled as following (Li et al., 2009)

\[
p(X_i|\Theta) = \sum_{j=1}^{M} \rho_{jd} \rho(X_i|\Theta) + (1 - \rho_{jd})p(X_i|\lambda_{jd}),
\]

where \( \Theta = \{\Theta_{0k}, \rho_{kd}, \lambda_{jd}\} \), \( \rho_{jd} \) denotes the weight of the \( d \)th feature on cluster \( j \) and \( p(X_i|\lambda_{jd}) \), \( \lambda_{jd} = (\lambda_{jkd}, \beta_{jkd}) \), is the Gamma distribution from which the feature is drawn if it is irrelevant. Notice that the previous model is reduced to the one in Eq. (3) when \( \rho_{jd} = \rho_{jd} \) and \( \lambda_{jd} = \lambda_{jd} \), \( j = 1, \ldots, M \), \( d = 1, \ldots, D \).

Generally no knowledge is available as to which vector \( X_i \) is not representative and then is not really generated from our assumed statistical model. Outliers can not only make the model learning more complex, but also corrupt the parameters estimates and then compromise the performance of the final model. Classic outliers identification methods have been generally based on the sample mean and covariance matrix which are themselves compromised by the outliers as shown in Rousseeuw and Van Zomeren (1990) especially in the case of high-dimensional non-Gaussian data. A better technique is to approach the problem by incorporating an auxiliary outlier component, to which we associate a uniform density\(^7\) Titsias and Williams (2003), Ke and Kanade (2004), Black and Jeppson (1996), and Williams and Titsias (2003), into the model:

\[
p(X_i|\Theta) = \sum_{j=1}^{M} \rho_{jd} \rho(X_i|\Theta) + (1 - \rho_{jd})p(X_i|\lambda_{jd}) + \rho_{M+1} U(X_i),
\]

where \( \rho_{M+1} = 1 - \sum_{j=1}^{M} \rho_{jd} \) is the probability that \( X_i \) was not generated by the central mixture model and \( U(X_i) \) 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 robustly unsupervised feature selection to learn the right meaning from the right observations (i.e. inliers). Notice that when \( \rho_{M+1} = 0 \) the outlier component is removed and the previous equation is reduced to Eq. (4).

3. Model learning

3.1. Maximum likelihood estimation

A well-known approach for unknown parameters estimation is the technique of maximum likelihood (ML) which properties have been extensively examined in the past (see, for instance, Hartley (1958)) using ML the parameters are estimated by maximizing the log-likelihood function as following

\[
\hat{\Theta} = \arg \max_{\Theta} \left\{ \log p(X|\Theta) = \sum_{i=1}^{N} \log \left[ \sum_{j=1}^{M} \left( \rho_{jd} \rho(X_i|\Theta) + (1 - \rho_{jd})p(X_i|\lambda_{jd}) \right) \right] + \rho_{M+1} U(X_i) \right\},
\]

which gives us (see Appendices A, B and C)

\[
p_j = \frac{\sum_{i=1}^{N} \rho_{jd} \rho(X_i|\Theta)}{N} \quad j = 1, \ldots, M + 1,
\]

\[
\rho_{jd} = \frac{\sum_{i=1}^{N} \rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_{jd})}{\sum_{i=1}^{N} \rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_{jd})} \quad j = 1, \ldots, M \quad d = 1, \ldots, D,
\]

\[
\beta_{jkd} = \frac{\sum_{i=1}^{N} \rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_{jd})}{\sum_{i=1}^{N} \rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_{jd})} \quad j = 1, \ldots, M \quad d = 1, \ldots, D,
\]

where \( \Psi^{-1}() \) is the inverse digamma function,

\[
p(j|X_i) = \left\{ \begin{array}{ll}
\sum_{j=1}^{M} \rho_j \rho(X_i|\Theta) H(X_i|\lambda_j) & \text{if } j = 1, \ldots, M, \\
\rho_{M+1} U(X_i) & \text{if } j = M + 1
\end{array} \right.
\]

is the posterior probability that a vector \( X_i \) will be considered as an inlier and then assigned to a cluster \( j, j = 1, \ldots, M \) or an outlier and then affected to cluster \( M + 1 \) which allows to safeguard against erroneous feature selection, and

\[
f(\rho_{jd} | \rho_{jd}, \lambda_{jd}) = \frac{\rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_j)}{\rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_j) + (1 - \rho_{jd})p(X_i|\lambda_{jd})},
\]

\[
f(1 - \rho_{jd} | \rho_{jd}, \lambda_{jd}) = \frac{(1 - \rho_{jd})p(X_i|\lambda_{jd})}{\rho_{jd} \rho(X_i|\Theta) H(X_i|\lambda_j) + (1 - \rho_{jd})p(X_i|\lambda_{jd})}.
\]

It is noteworthy that \( p(j|X_i, f(\rho_{jd}, \lambda_{jd}), (1 - \rho_{jd}, \lambda_{jd})) \) can be viewed as the posterior probability that a given feature \( d \) is relevant (irrelevant) for a cluster \( j \).

3.2. Model selection based on the integrated likelihood

It is desirable to determine the simplest model that can explain the data accurately.\(^6\) It is noteworthy that the simplicity is

\[^{6}\] Generally it is easier to understand simple models which are often the best solutions in accordance with Occam's Razor philosophy.
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 integrated (or marginal) likelihood, to reach an acceptable balance between model complexity and goodness of fit, given as follows when approximated using Laplace’s method (McLachlan and Peel, 2000).

\[
\log p(\mathbf{X}) \simeq \log p(\mathbf{X}|\hat{\Theta}) + \log p(\hat{\Theta}) - \frac{1}{2} \log(|H(\hat{\Theta})|) + \frac{N_p}{2} \log(2\pi),
\]

(16)

where \(\hat{\Theta}\) is the posterior mode, \(p(\Theta)\) is the prior density for \(\Theta\), \(H(\hat{\Theta})\) is the negative Hessian matrix evaluated at \(\hat{\Theta}\) and \(|H(\hat{\Theta})|\) is its determinant, and \(N_p\) is the number of free parameters to be estimated which is equal to \(M(5D + 1)\) in our case. More details and discussions about the integrated likelihood can be found in McLachlan and Peel (2000), Chickering and Heckerman (1997) and Bouguila and Ziou (2007) and references therein. In the following we develop the required terms to determine the integrated likelihood, of our model, namely the prior density and the determinant of the Hessian matrix.

Concerning the prior density \(p(\Theta)\), in the absence of knowledge about the parameters (or complete ignorance), a widely applied reasonable assumption is to consider that the mixture parameters are independent:

\[
p(\Theta) = p(p_1, \ldots, p_{M+1}) \prod_{j=1}^{M} \prod_{d=1}^{D} p(\rho_{jd}) p(\beta_{jd}) p(x_{jd})p(\alpha_d). \quad (17)
\]

It is common to assume a symmetric Dirichlet distribution with parameters \(\eta\) as a prior for the mixing parameters, since they are defined on the simplex \(p_1, \ldots, p_{M+1}; \sum_{i=1}^{M+1} p_i < 1\). Generally \(\eta\) is set to 1 which gives the following prior \(p(p_1, \ldots, p_{M+1}) = M!\). We know also that each \(\rho_{jd}\) is defined in the compact support \([0,1]\), thus a common widely used prior is the Beta distribution. Taking a symmetric Beta with parameters set to 1 gives us a uniform prior \(p(\rho_{jd}) = U_{[0,1]}\). For the scale parameters \(p(\beta_{jd})\) and \(\beta_{ijd}\), we consider \(p(\beta_{jd}) = \frac{1}{\beta_{jd}}\) and \(p(\beta_{ijd}) = \frac{1}{\beta_{ijd}}\) as priors, respectively. Moreover, we consider exponential priors, with parameters set to \(10^{-2}\), for the shape parameters: \(p(x_{jd}) = 10^{-2\exp(-10^{-2}x_{jd})}\) and \(p(x_{ijd}) = 10^{-2\exp(-10^{-2}x_{ijd})}\). Notice that these specific choices of priors express actually our uncertainty about the model’s parameters and were found convenient according to our experiments. By substituting all these priors into Eq. (17) we obtain

\[
p(\Theta) = M! \prod_{j=1}^{M} \prod_{d=1}^{D} 10^{-4} \frac{\exp \left( -10^{-2}(x_{jd} + x_{ijd}) \right)}{\beta_{jd} \beta_{ijd}}. \quad (18)
\]

Concerning the determinant of the Hessian matrix, it is common to assume, in the case of finite mixture models, that the components decouple (Roberts & Rezek, 1998), thus we may write \(|H(\Theta)|\) as

\[
|H(\Theta)| = |H(p_1, \ldots, p_{M+1})| \prod_{j=1}^{M} |H(\rho_{jd})| |H(\beta_{jd})|, \quad (19)
\]

where \(|H(p_1, \ldots, p_{M+1})|, |H(\rho_{jd})|, |H(\beta_{jd})|\) and \(|H(\rho_{jd})|\) are the determinants of the Hessians with respect to the mixing parameters, \(\theta_j\), \(\lambda_j\) and \(\rho_{jd}\), respectively, and are given by (see Appendix D)

\[
|H(p_{1, \ldots, p_{M+1}})| = \prod_{j=1}^{M} \prod_{i=1}^{N} \left( \frac{p(M+1)_{ji}X_i - p(j)_{X_i}}{p_j} \right)^2, \quad (20)
\]

\[
|H(\rho_{jd})| = \sum_{i=1}^{N} p(j) \frac{(1 - \rho_{jd})}{\rho_{jd}} \frac{f(1 - \rho_{jd}, \theta_j, \lambda_j)}{f(\rho_{jd}, \theta_j, \lambda_j)}^2, \quad (21)
\]

\[
|H(\beta_{jd})| = \sum_{i=1}^{N} p(j) \frac{(1 - \rho_{jd})}{\rho_{jd}} \frac{f(1 - \rho_{jd}, \theta_j, \lambda_j)}{f(\rho_{jd}, \theta_j, \lambda_j)}^2, \quad (22)
\]

\[
|H(\theta_j)| = \prod_{d=1}^{D} \left( \sum_{i=1}^{N} \frac{p(j) X_i f(\rho_{jd}, \theta_j, \lambda_j)}{\rho_{jd}^2} \left( \frac{X_i - \bar{X}_d}{\bar{y}_{jd}} \right) \right) \times \Psi_1(\beta_{jd}) \left( \sum_{i=1}^{N} p(j) X_i f(\rho_{jd}, \theta_j, \lambda_j) \right)^2, \quad (23)
\]

Having all the estimation equations and the integrated likelihood expression in hand, our model learning will be performed under the standard two-phase paradigm employed by the expectation maximization (EM) framework as follows:

Algorithm

For each candidate value of \(M\):

1. Set \(\rho_{jd} = 0.5, d = 1, \ldots, D, j = 1, \ldots, M\) and initialization of the rest of parameters.⁉
2. Iterate the two following steps until convergence:
   (a) E-step: Update \(p(j|X_i)\) using Eq. (13) and \(r(j|\rho_{jd}, \theta_j, \lambda_j)\) using Eq. (14)
   (b) M-step: Update the \(p_{jd}, \rho_{jd}, \beta_{jd}, \alpha_d, \beta_{ijd}\) and \(\lambda_{ijd}\) using Eqs. (7)-(12), respectively.
3. Calculate the associated integrated likelihood using Eq. (16).
4. Select the optimal model that yields the largest integrated likelihood.

The previous algorithm is based on the EM approach and both E- and M-steps have a complexity of \(O(NMD)\). According to our operational definition of outliers, they should have a uniform distribution, since they do not follow the pattern of the majority of the data. A common approach, to define this uniform distribution, is to suppose that the data follow a single component model averaged over all the observation (Titias & Williams, 2003). Thus, in our case, we choose the following⁺¹ \(U(X) = 1/N \sum_{i=1}^{N} \sum_{d=1}^{D} \left( \rho_{jd} p(X_i|\theta_j) + (1 - \rho_{jd}) p(X_i|\lambda_j) \right)\), where the parameters \(\rho, \theta_j\) and \(\lambda_j\) are estimated using ML technique, which takes into account the fact that outliers should be sparsely distributed. It is noteworthy that the previous algorithm allows to first detect outlying data. Then, the remaining ones (i.e. inliers) are used to identify the optimal clustering structure in terms of number of clusters, relevant features and optimal parameters.

4. Experimental results

In this section some experiments are carried out to evaluate the usefulness of our model. The experiments are performed on both synthetic and real data. Moreover, we investigate our approach on a challenging problem namely objects shape clustering. For purpose of comparisons, we have implemented our approach with

---

⁉ 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.

¹ The other choices are possible, but in our case this specific choice was found appropriate according to our experimental results.
Table 1
Parameters used to generate the synthetic data sets ($n_j$ represents the number of elements in cluster $j$).

<table>
<thead>
<tr>
<th>Data set</th>
<th>$j$</th>
<th>$x_{j1}$</th>
<th>$y_{j1}$</th>
<th>$x_{j2}$</th>
<th>$y_{j2}$</th>
<th>$n_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 1</td>
<td>1</td>
<td>18</td>
<td>4</td>
<td>13</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>Data set 2</td>
<td>1</td>
<td>18</td>
<td>4</td>
<td>13</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>38</td>
<td>7</td>
<td>39</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>Data set 3</td>
<td>1</td>
<td>18</td>
<td>4</td>
<td>13</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>38</td>
<td>7</td>
<td>39</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>49</td>
<td>9</td>
<td>46</td>
<td>8</td>
<td>100</td>
</tr>
</tbody>
</table>

Fig. 1. Integrated likelihood as a function of the number of clusters for the three synthetic data sets when relevancy of features is considered. Row 1: using localized feature selection, Row 2: using globalized feature selection.

Fig. 2. Localized features saliency for the synthetic data sets.
Gaussian mixture models, also. In all our experiments, we investigate the advantages of performing simultaneous outliers detection and feature selection. Moreover, all results are averaged over 10 runs of the algorithm. Due to space limitations, the results of comparing clustering with global and local feature selection are, however, reported only for the synthetic data.

4.1. Synthetic data

4.1.1. Experiment 1

The first application evaluates the performance of the proposed model using three 2D synthetic data sets generated from 2-3- and 4-components bivariate Gamma mixture models. The parameters used to generate these data sets are given in Table 1. The first experiment is conducted by appending eight “noisy” features to the generated data sets which increases the dimensionality of the data to 10. The goal of this experiment is to evaluate the ability of the algorithm in selecting features when no outliers are present. Fig. 1 shows the number of clusters selected for each data set using the integrated likelihood criterion when localized and globalized feature selection are considered. According to these figures, it is clear that our algorithm is able to select the correct number of clusters in both cases.

Table 2 contains the classification results for the three synthetic data sets using the integrated likelihood criterion when relevancy of features is considered and when all data are considered as inliers. Row 1: using localized feature selection, Row 2: using globalized feature selection.

![Fig. 3. Globalized features saliency for the synthetic data sets.](image)

![Fig. 4. Integrated likelihood as a function of the number of clusters for the three synthetic data sets when relevancy of features is considered and when all data are considered as inliers. Row 1: using localized feature selection, Row 2: using globalized feature selection.](image)
Fig. 5. Localized saliency of features for the synthetic data sets without outliers rejection.

Fig. 6. Globalized saliency of features for the synthetic data sets without outliers rejection.

Fig. 7. Integrated likelihood as a function of the number of clusters for the three synthetic data sets when relevancy of features is considered and outliers detection is performed. Row 1: using localized feature selection, Row 2: using globalized feature selection.
Fig. 8. Localized saliency of features for the synthetic data sets with outliers rejection.

Fig. 9. Globalized saliency of features for the synthetic data sets with outliers rejection.

Fig. 10. Sample images from the Vistex data set.

Table 3
Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions without taking into consideration the relevancy of features nor the presence of outlier data.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>183</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>198</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>143</td>
<td>8</td>
<td>0</td>
<td>49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>175</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>132</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>186</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>191</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>158</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>179</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>194</td>
<td>0</td>
</tr>
<tr>
<td>Outlier class</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
localized and globalized feature selection methods. The results in these table show that the localized feature selection approach improves significantly the classification accuracy, for the three generated data sets, compared to the globalized one.

4.1.2. Experiment 2

The second experiment is conducted by considering the presence of outlier data through the introduction of 5, 10 and 15 outlying 10-dimensional vectors into the first, second, and third generated data sets, respectively. Fig. 4 shows the number of clusters found using the integrated likelihood approach after performing feature selection by considering that all the vectors are actually inliers (i.e without performing outliers detection). According to this figure, the algorithm is unable to determine the exact number of clusters only for the third data set when performing globalized feature selection. Despite this fact, it is clear that the presence of outliers compromises the feature selection process by affecting high weights to some irrelevant features and by decreasing the relevancy weights of the two first features as shown in Figs. 5 and 6. When performing outliers detection, our algorithm is able to detect all the outlying data for the three generated data sets and to selected the exact number of clusters in all cases (see Fig. 7) which significantly improves the feature selection process as shown in Figs. 8 and 9.

4.2. Real data

The second application concerns the classification of handwritten numerals using a data set, composed of 10 classes ("0"–"9"), obtained from Blake and Merz (XXXX). Each class contains 200 patterns where each pattern is represented by a 47-dimensional positive vector describing extracted Zernike moments magnitudes features. We add to this data set 20 47-dimensional vectors, representing Zernike moments magnitudes features, extracted from 20 textural images from the MIT Vistex gray level texture database11

Table 4
Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions without taking into consideration the relevancy of features nor the presence of outlier data.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Outlier class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>152</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>187</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>42</td>
<td>147</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>197</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>125</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>46</td>
<td>151</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>174</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>182</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>153</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5
Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions by taking into consideration the relevancy of features and the presence of outlier data.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Outlier class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>186</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>192</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>179</td>
<td>0</td>
<td>3</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>198</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>182</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>184</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>194</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>181</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6
Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions by taking into consideration the relevancy of features and the presence of outlier data.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Outlier class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>161</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>189</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>129</td>
<td>0</td>
<td>0</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>172</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>185</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>162</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>194</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>182</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>187</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Outlier class | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |

Fig. 11. Localized saliency of features, for the handwritten numerals data set, using mixture of Gamma distributions when performing outliers detection.

Fig. 12. Localized saliency of features, for the handwritten numerals data set, using mixture of Gaussian distributions when performing outliers detection.

Fig. 13. Samples of the MPEG-7 CE Shape-1 Part-B data set.
These 20 added vectors are considered as our outlier data. The main goal of this application is to compare the performance of Gaussian and Gamma mixture models when dealing with features defined over the positive reals.

The first experiment is conducted to compare mixtures of Gamma and Gaussian distributions without taking into consideration the relevancy of features nor the presence of outlier data. Table 3 shows the confusion matrix when applying mixture of Gamma distributions; the number of miss-classified images is 281 which represents an accuracy of 86.08%. On the other hand, Table 4 shows the confusion matrix using mixture of Gaussian distributions; the number of miss-classified images is 389 which represents an accuracy of 80.74%. In both cases we are unable to detect the outlier data as a different class. Indeed, all outliers are affected to the 10 selected classes.

The second experiment is conducted using mixtures of Gamma and Gaussian distributions by taking into consideration the relevancy of features and by performing outliers detection. Note that only the results of localized feature selection are given, since we have found that it performs better than the globalized one. Table 5 shows the confusion matrix for the mixture of Gamma distributions; the number of miss-classified images is 104 which represents an accuracy of 94.85 percent. On the other hand, Table 6 shows the confusion matrix for mixture of Gaussian distributions; the number of miss-classified shapes is 241 which represents an accuracy of 88.06 percent. It is noteworthy that using our approach applied with the Gamma mixture we are able to detect all the outliers which is not the case with the Gaussian mixture (2 outliers are affected to class 9). Figs. 11 and 12 show the localized saliency of features for both mixtures.

4.3. 2-D Shape clustering

Shape modeling and representation is an important step in several applications such as content-based image retrieval and indexing, and image segmentation (Peng, Kim, Lee, & Chung, 2010). Zernike Moments Magnitudes (ZMMs), which have been the subject of extensive theoretical and experimental research in the past (Kintner, 1976; Teh & Chin, 1988; Hosny, 2010), are known to be very effective to model objects shapes. Indeed, they allow invariant (i.e. regardless the position, size and orientation) recognition of objects (Khotanzad & Hong, 1990). Moreover, the Gamma distribution was found suitable to model ZMMs (Kim & Kim, 1998). Thus, we present in the following results of applying our model on the clustering of shape images represented by ZMMs. It is noteworthy that the main goal here is to investigate our model in the case of shape clustering and not to propose a new approach for shape modeling. The data set used is a subset of the MPEG-7 CE Shape-1 Part-B data set that consists of seven classes, where each class includes 20 shape samples. Fig. 13 shows examples of

(see Fig. 10). These 20 added vectors are considered as our outlier data. The main goal of this application is to compare the performance of Gaussian and Gamma mixture models when dealing with features defined over the positive reals.

The first experiment is conducted to compare mixtures of Gamma and Gaussian distributions without taking into consideration the relevancy of features nor the presence of outlier data. Table 3 shows the confusion matrix when applying mixture of Gamma distributions; the number of miss-classified images is 281 which represents an accuracy of 86.08%. On the other hand, Table 4 shows the confusion matrix using mixture of Gaussian distributions; the number of miss-classified images is 389 which represents an accuracy of 80.74%. In both cases we are unable to detect the outlier data as a different class. Indeed, all outliers are affected to the 10 selected classes.

The second experiment is conducted using mixtures of Gamma and Gaussian distributions by taking into consideration the relevancy of features and by performing outliers detection. Note that only the results of localized feature selection are given, since we have found that it performs better than the globalized one. Table 5 shows the confusion matrix for the mixture of Gamma distributions; the number of miss-classified images is 104 which represents an accuracy of 94.85 percent. On the other hand, Table 6 shows the confusion matrix for mixture of Gaussian distributions; the number of miss-classified shapes is 241 which represents an accuracy of 88.06 percent. It is noteworthy that using our approach applied with the Gamma mixture we are able to detect all the outliers which is not the case with the Gaussian mixture (2 outliers are affected to class 9). Figs. 11 and 12 show the localized saliency of features for both mixtures.

4.3. 2-D Shape clustering

Shape modeling and representation is an important step in several applications such as content-based image retrieval and indexing, and image segmentation (Peng, Kim, Lee, & Chung, 2010). Zernike Moments Magnitudes (ZMMs), which have been the subject of extensive theoretical and experimental research in the past (Kintner, 1976; Teh & Chin, 1988; Hosny, 2010), are known to be very effective to model objects shapes. Indeed, they allow invariant (i.e. regardless the position, size and orientation) recognition of objects (Khotanzad & Hong, 1990). Moreover, the Gamma distribution was found suitable to model ZMMs (Kim & Kim, 1998). Thus, we present in the following results of applying our model on the clustering of shape images represented by ZMMs. It is noteworthy that the main goal here is to investigate our model in the case of shape clustering and not to propose a new approach for shape modeling. The data set used is a subset of the MPEG-7 CE Shape-1 Part-B data set that consists of seven classes, where each class includes 20 shape samples. Fig. 13 shows examples of

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Confusion matrix for MPEG-7 data set using mixture of Gamma distributions without taking into consideration the relevancy of features nor the presence of outlier data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified as</td>
<td>Bone</td>
</tr>
<tr>
<td>Bone</td>
<td>17</td>
</tr>
<tr>
<td>Heart</td>
<td>0</td>
</tr>
<tr>
<td>Glass</td>
<td>0</td>
</tr>
<tr>
<td>Fountain</td>
<td>0</td>
</tr>
<tr>
<td>Key</td>
<td>0</td>
</tr>
<tr>
<td>Fork</td>
<td>0</td>
</tr>
<tr>
<td>Hummer</td>
<td>4</td>
</tr>
<tr>
<td>Outlier class</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions without taking into consideration the relevancy of features nor the presence of outlier data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified as</td>
<td>Bone</td>
</tr>
<tr>
<td>Bone</td>
<td>19</td>
</tr>
<tr>
<td>Heart</td>
<td>4</td>
</tr>
<tr>
<td>Glass</td>
<td>0</td>
</tr>
<tr>
<td>Fountain</td>
<td>0</td>
</tr>
<tr>
<td>Key</td>
<td>5</td>
</tr>
<tr>
<td>Fork</td>
<td>1</td>
</tr>
<tr>
<td>Hummer</td>
<td>12</td>
</tr>
<tr>
<td>Outlier class</td>
<td>3</td>
</tr>
</tbody>
</table>

Fig. 14. Localized saliency of features using mixture of Gamma distributions without outliers detection.

12 http://www.cis.temple.edu/latecki/TestData/mpeg7shapeB.tar.gz.
images from this data set. We also add to this data set 5 textural images from the MIT Vistex gray level database as outliers. After normalizing all the images (Mukundan & Ramakrishnan, 1998; Abu-Mostafa & Psaltis, 1985), the vector of characteristics (ZMMs) is computed for each image using the method proposed in Kim and Kim (1998). Thus, each image is represented by a 36-dimensional vector.

The first experiment is conducted using mixtures of Gamma and Gaussian distributions without performing nor feature selection neither outliers detection. Table 7 shows the confusion matrix for mixture of Gamma; the number of miss-classified shapes is 29 which represents an accuracy of 80.00 percent. On the other hand, Table 8 shows the confusion matrix for mixture of Gaussian distributions; the number of miss-classified shapes is 50 which represents an accuracy of 65.51 percent. For both mixtures, the number of clusters selected by the integrated likelihood criterion is 7.

The second experiment is conducted using mixtures of Gamma and Gaussian distributions by performing feature selection without outliers detection. Tables 9 and 10 show the confusion matrices for both mixtures in this case. The number of miss-classified shapes using mixture of Gamma is 22 which represents an accuracy of 84.82 percent. On the other hand, the number of miss-classified shapes using mixture of Gaussian is 40 which represents an accuracy of 72.41 percent. Figs. 14 and 15 show the localized saliency of features of the classified shapes using both mixtures. Comparing

---

**Table 9**

Confusion matrix for MPEG-7 data set using mixture of Gamma distributions that takes into consideration the relevancy of features without outliers detection.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>Bone</th>
<th>Heart</th>
<th>Glass</th>
<th>Fountain</th>
<th>Key</th>
<th>Fork</th>
<th>Hummer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heart</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Glass</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fountain</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Key</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Fork</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Hummer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Outlier class</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**Table 10**

Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions that takes into consideration the relevancy of features without outliers detection.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>Bone</th>
<th>Heart</th>
<th>Glass</th>
<th>Fountain</th>
<th>Key</th>
<th>Fork</th>
<th>Hummer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Heart</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Glass</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Fountain</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Key</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fork</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Hummer</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Outlier class</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Table 11**

Confusion matrix for MPEG-7 data set using mixture of Gamma distributions by taking into consideration the relevancy of features and the presence of outlier data.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>Bone</th>
<th>Heart</th>
<th>Glass</th>
<th>Fountain</th>
<th>Key</th>
<th>Fork</th>
<th>Hummer</th>
<th>Outlier class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Heart</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Glass</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fountain</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Key</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fork</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Hummer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Outlier class</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 12
Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions by taking into consideration the relevancy of features and the presence of outlier data.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>Bone</th>
<th>Heart</th>
<th>Glass</th>
<th>Fountain</th>
<th>Key</th>
<th>Fork</th>
<th>Hummer</th>
<th>Outlier class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Heart</td>
<td>0</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Glass</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fountain</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Key</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Fork</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hummer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>Outlier class</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Fig. 16. Localized saliency of features using mixture of Gamma distributions by taking into consideration the presence of outlier data.

Fig. 17. Localized saliency of features using mixture of Gaussian distributions by taking into consideration the presence of outlier data.
the first and second experiments, we can deduce that the feature selection process improves significantly the clustering of shapes accuracy. However, as 7 is considered as the optimal number of clusters, for both mixtures, the outliers are mixed with the inliers and affected to these clusters. The third experiment is conducted by taking into consideration both the relevancy of features and the presence of outlier data (in the case of the Gaussian, however, one outlier was considered as an inlier); hence, the classification accuracy has increased. Tables 11 and 12 show the confusion matrices for both mixtures. The numbers of miss-classified shapes are 9 and 21 which represent accuracies of 93.79 and 85.51 percent when using Gamma and Gaussian mixtures, respectively. Figs. 16 and 17 show the localized saliency of features for both mixtures.

5. Conclusion

In this paper, we have presented a principled unsupervised generative model-based approach to simultaneous clustering, feature selection and outlier detection for the purpose of robust data modeling using finite multi-dimensional Gamma mixture models. Our proposal was to use a statistical model that makes explicit what data or features have to be ignored and what information has to be retained. Our work is mainly driven by the increased collection of high-dimensional non-Gaussian data in various domains and by the complexity of both feature selection and outlier detection problems in such domains. It has been shown through extensive experiments involving synthetic and real data that the proposed approach has excellent modeling capabilities and that feature selection mixed with outliers detection influences significantly the clustering performance. Future works can be devoted to extend the proposed model to online settings using a variational approach for instance, since we generally deal with dynamically changing environments. The proposed model can be extended also to the semi-supervised case.

Acknowledgment

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

Appendix A. Proof of Eq. (7)

Note that we have to introduce a lagrange multiplier A to incorporate the constraint \( \sum_{j=1}^{M+1} p_j = 1 \). Computing the derivative of \( \log p(X|\Theta) \) w.r.t. \( p_j, j = 1, \ldots, M+1, \) we obtain
\[
\frac{\partial}{\partial p_j} \left[ \log \left( \sum_{j=1}^{M+1} p_j \prod_{i=1}^{K} p_j(X_i|\theta_j) \right) + \left( 1 - \rho_j \right) p_j(X_i|\theta_j) \right] = \frac{1}{p_j} \sum_{i=1}^{M+1} p_j(X_i|\theta_j) - A = 0.
\]
\[
\chi_{\theta_j} = \Psi^{-1} \left( \frac{\sum_{i=1}^{M+1} p_j(X_i|\theta_j) \prod_{i=1}^{K} p_j(X_i|\theta_j) \log p(X_i|\theta_j)}{\sum_{i=1}^{M+1} p_j(X_i|\theta_j) \prod_{i=1}^{K} p_j(X_i|\theta_j) \log p(X_i|\theta_j) - \rho_j p(X_i|\theta_j) \log p(X_i|\theta_j)} \right),
\]
where \( \Psi^{-1}() \) is the inverse digamma function.

Appendix B. Proof of Eq. (8)

Let \( \rho_{d,1} = \rho_{d,1} = 1 - \rho_d = \rho_{d,2} \). By computing the derivative w.r.t. \( \rho_{d,1}, j = 1, \ldots, M, d = 1, \ldots, D \) and introducing the lagrange multiplier \( A \) to take into account the fact that \( \rho_{d,1} + \rho_{d,2} = 1 \), we obtain
\[
\frac{1}{\partial \rho_{d,1}} \sum_{j=1}^{M+1} \log \left( \prod_{i=1}^{K} p_j(X_i|\theta_j) \left( \rho_{d,1} p_j(X_i|\theta_j) + \rho_{d,2} p_j(X_i|\theta_j) \right) \right) + p_{M+1} U(X_i) = A(1 - \rho_{d,1} - \rho_{d,2}) = \sum_{i=1}^{N} p(j|\theta_i) \left( \rho_{d,1} p(X_i|\theta_j) + \rho_{d,2} p(X_i|\theta_j) \right) - A = 0.
\]
\[
\rho_{d,1} = \sum_{i=1}^{N} p(j|\theta_i) \left( \frac{\rho_{d,1} p(X_i|\theta_j)}{\rho_{d,1} p(X_i|\theta_j) + \rho_{d,2} p(X_i|\theta_j)} \right) - A = 0.
\]
\[
\rho_{d,2} = \sum_{i=1}^{N} p(j|\theta_i) \left( \frac{\rho_{d,2} p(X_i|\theta_j)}{\rho_{d,1} p(X_i|\theta_j) + \rho_{d,2} p(X_i|\theta_j)} \right) - A = 0.
\]
By computing the derivative w.r.t. \( \rho_{d,1} \) we obtain
\[
\sum_{i=1}^{N} p(j|\theta_i) \frac{p_{d,1} p(X_i|\theta_j)}{p_{d,1} p(X_i|\theta_j) + p_{d,2} p(X_i|\theta_j)} - A = 0.
\]
\[
\rho_{d,2} = \sum_{i=1}^{N} p(j|\theta_i) \left( \frac{\rho_{d,2} p(X_i|\theta_j)}{\rho_{d,1} p(X_i|\theta_j) + \rho_{d,2} p(X_i|\theta_j)} \right) - A = 0.
\]

Appendix C. Proof of Eqs. (9) and (10)

Computing the derivative of \( \log p(X|\Theta) \) w.r.t. \( \beta_{d,1} \), we obtain
\[
\frac{\partial \log p(X|\Theta)}{\partial \beta_{d,1}} = \sum_{i=1}^{N} p(j|\theta_i) \frac{\partial}{\partial \beta_{d,1}} \log \left( \rho_{d,1} p(X_i|\theta_j) + (1 - \rho_{d,1}) p(X_i|\theta_j) \right) = \sum_{i=1}^{N} p(j|\theta_i) \frac{\partial}{\partial \beta_{d,1}} \log p(X_i|\theta_j),
\]
where
\[
p(j|\theta_i) = \frac{\rho_{d,1} p(X_i|\theta_j)}{\rho_{d,1} p(X_i|\theta_j) + (1 - \rho_{d,1}) p(X_i|\theta_j)}.
\]
we have
\[
\frac{\partial}{\partial \beta_j} \log p(X_{ij} | \theta) = \frac{\partial}{\partial \beta_j} \left( (x_{ij} - 1) \log X_{ij} - \beta_j \log \beta_j - \log \Gamma(\beta_j) \right)
\]
\[
= X_{ij} \beta_j - \beta_j \log \beta_j - \log \Gamma(\beta_j)
\]
Then,
\[
\frac{\partial \log p(X|\Theta)}{\partial \beta_j} = 0 \text{ gives us} \beta_j = \frac{\sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j)}{\sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j)}
\]
Computing the derivative of \( \log p(X|\Theta) \) w.r.t \( \beta_j \), we obtain
\[
\frac{\partial \log p(X|\Theta)}{\partial \beta_j} = \sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j) - (1 - \rho_j) p(X_{ij}|\theta_j)
\]
we have
\[
\frac{\partial}{\partial x_{ij}} \log p(X_{ij} | \theta_j) = \frac{\partial}{\partial x_{ij}} \left( (x_{ij} - 1) \log X_{ij} - \beta_j \log \beta_j - \log \Gamma(\beta_j) \right)
\]
\[
= \log X_{ij} - \log \beta_j - \Psi(\beta_j),
\]
where \( \Psi() \) is the digamma function. Thus
\[
x_{ij} = x_{ij}^{\text{new}} - \frac{\partial \log p(X_{ij} | \theta_j)}{\partial x_{ij}} \left( \frac{\partial \log p(X|\Theta)}{\partial \beta_j} \right)^{-1}
\]
It is straightforward to determine \( x_{ij}^{\text{new}} \) and \( \beta_{ij} \) following the same development as above.

Appendix D. Proof of Eqs. (20)–(23)

The Hessian matrix is defined as, \( H(\theta) = \frac{\partial^2 \log p(X|\Theta)}{\partial \theta^2} \). Its determinant can be defined as follows: \( |H(\theta)| = |H(p_1, \ldots, p_{M+1})| \prod_{i=1}^{M+1} \left| \frac{\partial}{\partial \theta} \log p(X_{ij} | \theta_j) \right| \prod_{i=1}^{M+1} \left| H(\theta_j) \right| \). Let's start by the Hessian matrix with respect to the mixing parameters \( |H(p_1, \ldots, p_{M+1})| \) which should take into account the fact that \( p_{M+1} = 1 - \sum_{j=1}^{M} p_j \). We have
\[
\frac{\partial^2 \log p(X|\Theta)}{\partial \theta^2} = \left| \begin{array}{cccc}
\sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j) & \cdots & \sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j) \\
\vdots & \ddots & \vdots \\
\sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j) & \cdots & \sum_{i=1}^{N} p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j)
\end{array} \right| - U(X_{ij})
\]
By approximating the Hessian, with respect to the mixing parameters, from its diagonal components only, it is possible to show that (Roberts & Rezek, 1998)
\[
|H(p_1, \ldots, p_{M+1})| = \prod_{j=1}^{M} \sum_{i=1}^{N} \frac{p(M+1|X_{ij}) - p(j|X_{ij})^2}{p_j} \cdot \frac{p(M+1|X_{ij}) - p(j|X_{ij})^2}{p_j}
\]
As for the Hessian, with respect to the \( \rho_j \) parameters, we have
\[
\frac{\partial}{\partial \rho_j} \sum_{i=1}^{N} \log \left( p(j|X_{ij}) \rho_j p(X_{ij}|\theta_j) \right) + \sum_{i=1}^{N} U(X_{ij})
\]
\[
= \sum_{i=1}^{N} p(j|X_{ij}) \frac{p(X_{ij}|\theta_j) - p(X_{ij}|\theta_j)}{\rho_j}
\]
\[
= \sum_{i=1}^{N} p(j|X_{ij}) \left( f(1 - \rho_j p(X_{ij}|\theta_j)) - f(\rho_j p(X_{ij}|\theta_j)) \right)
\]
An the determinant of the Hessian can be approximated as
\[
|H(\rho_j)| = \sum_{i=1}^{N} p(\rho_j|X_{ij})^2 \left( f(1 - \rho_j p(X_{ij}|\theta_j)) - f(\rho_j p(X_{ij}|\theta_j)) \right)^2.
\]
According to the previous appendix, it is easy to show that:
\[
\frac{\partial^2 \log p(X|\Theta)}{\partial \theta^2} = - \sum_{i=1}^{N} p(j|X_{ij}) f(\rho_j, \theta_j, \lambda_j)
\]
\[
\frac{\partial^2 \log p(X|\Theta)}{\partial \theta^2} = \Psi(\theta_j) \sum_{i=1}^{N} p(j|X_{ij}) f(\rho_j, \theta_j, \lambda_j)
\]
References


