A possibilistic clustering approach toward generative mixture models

Sotirios P. Chatzis*, Gavriil Tsechpenakis

Imperial College London, Department of Electrical and Electronic Engineering, Exhibition Road, London SW7 2BT, United Kingdom

**Article Info**

**ABSTRACT**

Generative mixture models (MMs) provide one of the most popular methodologies for unsupervised data clustering. MMs are formulated on the basis of the assumption that each observation derives from (belongs to) a single cluster. However, in many applications, data may intuitively belong to multiple classes, thus rendering the single-cluster assignment assumptions of MMs irrelevant. Furthermore, even in applications where a single-cluster data assignment is required, the induced multinomial allocation of the modeled data points to the clusters derived by a MM, imposing the constraint that the membership probabilities of a data point across clusters sum to one, makes MMs very vulnerable to the presence of outliers in the clustered data sets, and renders them ineffective in discriminating between cases of equal evidence or ignorance. To resolve these issues, in this paper we introduce a possibilistic formulation of MMs. Possibilistic clustering is a methodology that yields possibilistic data partitions, with the obtained membership values being interpreted as degrees of possibility (compatibilities) of the data points with respect to the various clusters. We provide an efficient maximum-likelihood fitting algorithm for the proposed model, and we conduct an objective evaluation of its efficacy using benchmark data.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Generative mixture models (MMs) have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena. Because of their usefulness as an extremely flexible method of modeling, MMs continued to receive increasing attention over the years, from both a practical and a theoretical point of view [1]. MMs consider that each data point is generated by one specific model component distribution, and, hence allocated to a single cluster. However, in many applications, a data point might be viewed as belonging to multiple (overlapping) classes, thus requiring the introduction of a less rigid clustering scheme in the context of MMs. This could be the case, e.g., in topic-based image retrieval, where typically one image might be related to multiple topics.

Relaxing this rigidity has constituted in the past a domain of research in the framework of cluster analysis. Many authors have proposed a fuzzy setting as the appropriate approach to cope with this problem. Fuzzy set theory, proposed by Zadeh [2] in 1965, gives an idea of uncertainty of belonging, which is described by a membership function. One of the most widely used approaches to fuzzy prototype-based clustering is fuzzy c-means (FCM) clustering, introducing the notion of the degree of fuzziness [3].

FCM-type fuzzy clustering algorithms have been shown to be closely related to MMs in the algorithmic framework. Hathaway [4] interpreted the expectation-maximization (EM) algorithm for mixtures of Gaussians (MoGs) as a penalized version of the hard clustering algorithm. Gan et al. [5] showed that the MoG can be translated to an additive fuzzy system. Ichihashi et al. [6] showed that the EM algorithm for the MoG can be derived from the FCM-type fuzzy clustering, when considering a regularized fuzzy objective function, for a proper selection of the distance metric. Recently, an FCM-type fuzzy clustering regard toward generative mixture models (of any form of component distributions), as well as hidden Markov random field models, along with the appropriate model fitting algorithms, were introduced in [7–10], being shown to provide significant advantages compared with conventional maximum-likelihood treatments of these models, using the EM algorithm.

A key characteristic of fuzzy clustering algorithms comprises their constraint that the cluster memberships of a given data point must sum up to one over the postulated clusters. In essence, this constraint implies a probabilistic clustering notion, giving rise to a clustering scheme which comprises a generalization of the multinomial data allocation scheme obtained by common MM formulations. As a direct implication of this construction, the membership of a data point to a cluster obtained by a fuzzy clustering algorithm is contingent on the memberships to all the other clusters; under a geometrical regard, this means that cluster membership functions depend not only on where the patterns are located with respect to the considered cluster, but also to the other clusters. Hence, in the...
framework of fuzzy clustering, the membership functions take the meaning of relative degrees of sharing, measuring the similarity of a pattern to a cluster relatively to the others, similar to the case of common MM formulations where the derived probabilistic membership functions are usually referred to as the data responsibilities (posterior probabilities of cluster membership).

Nevertheless, this relativity of the nature of the obtained membership functions can be a major disadvantage of the clustering algorithm and gives rise to a number of problems [11]. Indeed, such constrained membership functions are not capable of distinguishing between equal evidence or ignorance; in other words, the obtained memberships are incapable of detecting cases of either data points which would be equally likely to belong to more than one clusters, or of data points that would be unlikely to belong to anyone of the known clusters, usually referred to as outliers [9].

Even worse, this incapacity of the obtained membership functions to pinpoint the outlying data might result in a severe under-mining of the performance of the clustering algorithm, resulting in a rather poor partition of the data [1]. Thus, it is not surprising that a great deal of research work has been devoted to the attenuation of these problems, both in the framework of generative mixture models [12,13,8], and the fuzzy clustering framework [10,9,14,15].

Possibilistic clustering has been proposed as an effective approach to overcome the above limitations, by recasting the clustering problem into the framework of possibility theory [13,16]. Under the possibilistic regard, the membership functions obtained by the clustering algorithm are interpreted as degrees of compatibility or possibility. This way, any number of clusters that a considered data point is representative of yield high membership values, while untypical points (outliers) bear low memberships to all clusters. Commencing with the seminal work of Krishnapuram and Keller [17], yielding the possibilistic c-means (PCM) algorithm, a number of possibilistic clustering algorithms have been derived in the recent years, and their advantages have been experimentally illustrated [14,15].

Motivated by the aforementioned merits of possibilistic clustering methodologies, in this paper we introduce a novel possibilistic approach toward MMVs. We begin with the assumption that the clusters to be derived can be modeled by a prespecified form of probability density functions (e.g., Gaussians). This way, a generative mixture model is in principle obtained. Further, we consider that each one of the analyzed data points may simultaneously be compatible with more than one of the postulated clusters that comprise the model. Under this regard, we seek to derive a proper possibilistic construction for the postulated MM. This way, the possibilistic mixture model (PMM) is derived. Our model can be seen as a continuous latent variable relaxation of clustering with finite mixture models. Unlike latent Dirichlet allocation (LDA) [18], and mixed membership models [1], which also capture partial membership in the form of attribute-specific mixtures, our model does not assume a factorization over attributes and provides a general way of combining Gaussian distributions with partial membership; this way, we obtain a soft allocation scheme of the observed data points over the clusters derived by the postulated mixture models.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of existing formulations of generative mixture models, and a brief review of possibilistic clustering algorithms. In Section 3, the proposed possibilistic approach to fuzzy c-means algorithms and its variants is presented. In Section 4, a computationally efficient maximum-likelihood (ML) fitting algorithm for the PMM is provided. In Section 5, we thoroughly evaluate the efficacy of the proposed model through a number of experiments, considering a series of demanding data clustering applications, using publicly available data. Finally, in the concluding section, we summarize and discuss our results.

2. Preliminaries

2.1. Generative mixture models

Let us consider a data set \( X = \{x_n\}_{n=1}^N \) of \( d \)-dimensional multivariate observations. Postulating a generative mixture model to represent this data set is a common approach in the field of statistical pattern recognition [19]. A \( K \)-component generative mixture model is, in essence, a superposition of weighted distributions, usually selected from the exponential family, of the form

\[
p(x_n) = \sum_{k=1}^{K} \pi_k f(x_n | \theta_k) \tag{1}
\]

where \( f(x_n | \theta_k) \) are the component distributions of the model, and \( \pi_k \) are their mixing weights (prior probabilities). As implied by (1), generative mixture models are based on the assumption that each modeled data point \( x_n \) is related with (generated by) one (and only one) of the mixture component distributions. Hence, under a data clustering viewpoint, MMVs can be regarded as providing a statistical framework for clustering, where each cluster is modeled by a postulated (model component) distribution, and each data point is eventually assigned to a single cluster [1]. Model training for MMVs can be conducted under either a maximum-likelihood setting, usually by means of the expectation-maximization algorithm and its variants [20], or under the Bayesian framework [19], using either stochastic Monte–Carlo techniques [21], or deterministic variational approximations [22].

2.2. Possibilistic clustering algorithms

As already discussed, most of the existing clustering algorithms, including MMVs and fuzzy clustering algorithms, impose a probabilistic constraint on the utilized membership functions that the obtained cluster memberships of a data point must sum up to one over the derived clusters. A direct repercussion of this construction is that such algorithms are not explicitly designed to effectively handle problems where a modeled data point might simultaneously belong to multiple clusters. Yet, even in cases where merely a single cluster assignment is asked for, this formulation of the clustering algorithm often results in a number of counterintuitive effects, at least in the sense of measuring the compatibilities of the modeled data points to the derived cluster prototypes. Consider, for example, an application of the clustering algorithm set to obtain two clusters, a data point \( A \) which is an equally “good” member of both clusters, and a data point \( B \) which is an equally “poor” member of both clusters. In that occasion, the probabilistic constraints imposed on the derived membership functions, giving rise to memberships providing only a relative measure of cluster compatibility for the modeled data points, would result in both the data points having memberships equal to 0.5 with respect to both the derived clusters. Obviously, this result is reasonable when regarding the computed membership functions as merely a means of obtaining a boundary separating the data into groups, still this setting is very different from what one would intuitively expect from a properly designed membership function, which should yield high values for the memberships of point \( A \) with respect to both clusters, and low membership values for point \( B \) with respect to both clusters.

Possibilistic clustering has been proposed as an effective way to ameliorate this counterintuitive behavior, and, hence, the induced vulnerability of conventional clustering algorithms to outliers. Possibilistic c-means was the first such algorithm [17], obtained by relaxation of the probabilistic constraints imposed on the membership function of the fuzzy c-means algorithm. More specifically,
the original FCM algorithm formulation minimizes the objective function given by [23]

\[ J_\phi = \sum_{k=1}^{K} \sum_{n=1}^{N} u_{nk}^\phi d_{nk} \]  

(2)

where \( d_{nk} \) is the dissimilarity (distance) functional of the \( n \)th data point with respect to the \( k \)th cluster, \( u_{nk} \) is the corresponding fuzzy membership function, \( \phi \) is the degree of fuzziness of the algorithm, and it is assumed that

\[ \sum_{k=1}^{K} u_{nk} = 1 \]  

(3)

Simply relaxing the probabilistic constraint in (3) would only produce the trivial solution, i.e. assign all the memberships to zero. For this reason, a new objective function is introduced in [17], yielding cluster compatibility functions (replacing the memberships of the FCM algorithm) which are as high as possible for the cluster yielding cluster compatibility functions (replacing the memberships of the original FCM algorithm formulation) that is the prior distribution of the allocation of the modeled data points, and low with respect to all the derived clusters. This way, the PC model is obtained, minimizing the objective function [17]

\[ J_\phi = \sum_{k=1}^{K} \sum_{n=1}^{N} u_{nk}^\phi d_{nk} + \sum_{k=1}^{K} \eta_k \sum_{n=1}^{N} (1-u_{nk})^\phi \]  

(4)

where \( \eta_k \) are suitable positive numbers and

\[ 0 < \sum_{n=1}^{N} u_{nk} < N, \quad \forall k \]  

(5)

Possibilistic clustering has attracted some significant research interest during the last years, with a number of extensions of the original FCM algorithm having been proposed in the relevant literature. For example, in [24], two interesting variations of PCM have been presented. Attempts to combine fuzzy and possibilistic clustering under a hybrid PCM-FCM formulation have been presented in [14] and [25]. In [26], a similarity-based PC algorithm is proposed, based on the idea to extend PCM for similarity-based clustering applications by integration with the mountain method [27].

Inspired by these advances, in this paper we introduce a possibilistic approach toward generative mixture models, providing a new regard of MMs as a statistical methodology for efficient possibilistic clustering. This way, we effectively introduce the merits of possibilistic clustering into the framework of statistical data modeling, giving rise to a novel, more flexible tool in the field of statistical pattern recognition. We shall introduce the proposed PMM in the following section.

3. Model formulation

Let us consider a data set \( X = \{x_n\}_{n=1}^{N} \) of \( d \)-dimensional multivariate observations. We want to obtain a possibilistic partition of this data set into \( K \) clusters. Let us also define a set of auxiliary variables \( Z = \{z_{nk}\}_{k=1}^{K} \) such that

\[ z_{nk} = \begin{cases} 1 & \text{if } x_n \text{ is viewed as generated from the } k \text{th cluster} \\ 0 & \text{otherwise} \end{cases} \]  

(6)

We will be referring to the \( z_{nk} \) variables as the cluster compatibility indicators; in the following we shall also use the notation \( z_{nk} = \{z_{nk}\}_{k=1}^{K} \).

Obviously, under our possibilistic clustering setting, the following may hold

\[ z_{nk} = 1 \land z_{n\ell} = 1, \quad \ell \neq k \]  

(7)

i.e., one data point may simultaneously be compatible with more than one cluster. Then, the prior probabilities of the \( z_{nk} \) variables, that is the prior distribution of the allocation of the modeled data to the clusters being derived, are easy to show that follow a Bernoulli distribution. Hence, using the definition

\[ \pi_k = P(z_{nk} = 1) \]  

(8)

we have that the \( z_{nk} \) are independent and identically distributed (i.i.d.) according to

\[ p(z_{nk} = \pi_k) = \text{Bernoulli}(\pi_k) \]  

(9)

That is

\[ p(z_{nk} | \pi_k) = \pi_k (1-\pi_k)^{1-z_{nk}} \]  

(10)

On the sequel, we take the conditional probabilities of the modeled data points, \( x_n \), given the cluster compatibility indicator variables \( z_{nk} \), to be i.i.d. according to an exponential family distribution \( f(\theta_k) \), such that

\[ p(x_n | z_{nk} = 1; \theta_k) = f(x_n | \theta_k) \]  

(11)

This way, a possibilistic analog to MMs is essentially obtained, yielding

\[ p(x_n) = \sum_{z_{nk}=1}^{N} P(z_{nk} | \pi_k) p(x_n | z_{nk}) = \sum_{z_{nk}=1}^{N} \prod_{k=1}^{K} \pi_{nk}^z(1-\pi_{nk})^{1-z_{nk}} f(x_n | \theta_k)^{z_{nk}} \]  

(12)

where \( \pi = \{\pi_{nk}\}_{n=1}^{N}, k \). We shall be denoting the derived model, with marginal distribution (12), as the possibilistic generative mixture model. It is interesting to compare the expression of the distribution \( p(x_n) \) for the introduced PMM with the corresponding expression for the case of a conventional (probabilistic) formulation of MMs. In that case, we would have [1]

\[ p(x_n) = \sum_{z_{nk}=1}^{N} p(z_{nk} | \pi) p(x_n | z_{nk}) = \sum_{z_{nk}=1}^{N} \prod_{k=1}^{K} \pi_{nk}^{z_{nk}} f(x_n | \theta_k)^{z_{nk}} \]  

(13)

with \( z_{nk} = 1 \land z_{n\ell} = 0 \forall \ell \neq k \). The comparison between (12) and (13) makes obvious that the introduced possibilistic formulation of the MM yields a model similar in structure to the conventional MM, but with a Binomial prior allocation \( p(z_{nk} | \pi) \) of the modeled data points to the derived clusters, contrary to the Multinomial prior allocation [19] implied by the conventional (probabilistic) MM formulation.

We would also like to underline that the general setting of our model is substantially different from that of popular statistical models which also capture partial membership, like latent Dirichlet allocation (LDA) [18]. Such models capture partial membership in the form of attribute-specific mixtures. For example, in LDA a document could be associated with multiple topics (clusters); however, this association is indirect, since LDA essentially considers the assignment of each one of the words comprising a modeled document to one specific (word-generating) topic (cluster); through this procedure, the modeled document is eventually associated with multiple topics (clusters), however, this is only an indirect assignment, in the form of an attribute-specific clustering mechanism. Similar, mixed membership (MM) models, or admixture models, assume that each data attribute (e.g. words) of the data point (e.g. document) is drawn independently from a mixture distribution given the membership vector for the data point, and hence, multiple-cluster assignment is an indirect and, in essence, “byproduct” procedure, based on a two-level organization of the modeled data. On the contrary, the proposed model does not assume a factorization over attributes but provides a general way of combining Gaussian distributions with partial membership; this way, we obtain a soft
allocation scheme of the observed data points directly over multiple clusters derived by the postulated mixture models.

4. Maximum-likelihood training for the PMM

To provide a model training algorithm for the PMM, we follow the ML paradigm deriving an efficient EM-type algorithm. Let us consider the PMM with marginal density function given by (12), and a d-dimensional fitting data set \( X = \{x_n\}_{n=1}^N \). To effect the EM-type treatment of the model, we regard the observed data \( X \) as being incomplete, as the cluster compatibility indicators \( \mathbf{z} \) are not available. The complete-data log likelihood is then obtained as

\[
\log L_c(\Psi) = \log \prod_{n=1}^N \prod_{k=1}^K \pi^c_{nk}(1 - \pi_{nk})^{1 - \log f(x_n | \theta_k)}
\]

whence, the E-step of the algorithm eventually reduces to the calculation of the compatibility posteriors \( q(z_{nk} = 1 | \Psi^{(t)}) \). To compute the compatibility posteriors \( q(z_{nk} = 1 | \Psi^{(t)}) \), we perform maximization of (18) over them by resorting to the simplex search method of Lagarias et al. [28].

4.2. M-step

To effect the M-step on the \((t + 1)\)th algorithm iteration, the objective function (16) of the EM algorithm for the PMM eventually yields

\[
Q(\Psi | \Psi^{(t)}) = \sum_{n=1}^N \sum_{k=1}^K q(z_{nk} = 1 | \Psi^{(t)}) \log \pi_{nk} + \log f(x_n | \theta_k)
+ \sum_{n=1}^N \sum_{k=1}^K (1 - q(z_{nk} = 1 | \Psi^{(t)}) \log (1 - \pi_{nk})
\]

Concerning the updates of \( \theta \) on the M-step on the \((t + 1)\)th iteration of the algorithm, it can be seen from (18) that \((\theta^{(t+1)})\) is obtained as an appropriate root of

\[
\sum_{n=1}^N \sum_{k=1}^K q(z_{nk} = 1 | \Psi^{(t)}) \log f(x_n | \theta_k) / \partial \theta_k = 0
\]

As we observe, one convenient feature of the EM algorithm for the PMM is that the solution of (20) often exists in closed form, as we shall specifically demonstrate for the case of Gaussian mixture component densities (Gaussian clusters selection), in the following section.

The E- and M-steps of the EM algorithm are alternated repeatedly until convergence of the objective function of the algorithm. The objective function of the EM algorithm is not decreased after one iteration [1]; further, under the very weak condition of the \( Q(\Psi | \Psi) \) being continuous over \( \Psi \) and \( \Psi \), the algorithm is guaranteed to converge to a local maximum (unless trapped to a saddle point, which, however, can be easily avoided by appropriate initialization) [29].

5. Illustration: the case of Gaussian mixture models

Typically, the component distributions of generative mixture models are taken to be of a multivariate Gaussian form. The popularity of this Gaussian selection stems from the simplicity of the updating equations of the resulting EM algorithm, as well as the well-known property of MoGs as a means to approximate unknown random distributions of any type, including distributions with multiple modes [1]. Motivated by these facts, in this section we provide the detailed derivations of the EM algorithm for the PMM with Gaussian component densities, which we shall denote as the possibilistic mixture of Gaussians (PMoG) model.

Let us postulate a PMM with Gaussian observation densities, that is with

\[
f(x_n | \theta_k) = N(x_n | \mu_k, \Sigma_k)
\]

We consider the EM algorithm for fitting this model to a set of training data \( X = \{x_n\}_{n=1}^N \). The E-step of the algorithm reduces to the computation of the estimates and the (approximate) posteriors of the compatibility indicator vectors, as described above. On the other hand, the M-step of the algorithm comprises the updates of the prior
probabilities, which can be obtained by (19), and the updates of the set of the internal parameters of the model component distributions, \(\{\theta_k\}_{k=1}^K = \{\mu_k, \Sigma_k\}_{k=1}^K\), which can be obtained by resolving Eq. (20). Using (21), it is easy to show that, on the \((t+1)\)th iteration of the EM algorithm for the PMoG, we have the updates

\[
\mu_k^{(t+1)} = \frac{\sum_{n=1}^N x_n q(z_{nk} = 1 | \Psi^{(t)}) \mu_k^{(t)} \Sigma_k^{-1} (x_n - \mu_k^{(t)})}{\sum_{n=1}^N q(z_{nk} = 1 | \Psi^{(t)})}
\]

\[
\Sigma_k^{(t+1)} = \frac{\sum_{n=1}^N q(z_{nk} = 1 | \Psi^{(t)}) (x_n - \mu_k^{(t+1)}) (x_n - \mu_k^{(t+1)})^T}{\sum_{n=1}^N q(z_{nk} = 1 | \Psi^{(t)})}
\]

6. Experiments

Here, we provide a thorough experimental evaluation of the PMM. In our investigations, we consider models with Gaussian observation densities (PMoG models), with full or diagonal covariance matrices. Our experiments are implemented in MATLAB, and are run on a Macintosh notebook, with an Intel Core 2 Duo 2 GHz CPU and 2 GB of RAM. In all applications, the performance of the PMoG model is compared to that of the MoG model, with the EM algorithm for both models being initialized at the same starting point. We postulate a MoG and a PMoG model comprising three observation densities (PMoG models), with full or diagonal covariances. In our investigations, we consider models with Gaussian distributions, and we use them to cluster the considered data. In Table 1, we provide the obtained clustering error rates for the considered simulated data points from the three Gaussians. We also provide the error rates obtained by the possibilistic fuzzy c-means algorithm (PFCM) of [14]. As we observe, the proposed approach outperforms both its alternatives in terms of the obtained clustering results.

- **Experiment on simulated data**

We begin with a toy example on simulated data, to demonstrate the notion behind the introduction of a Binomial data allocation scheme in the context of generative mixture models, and, hence, illustrate the advantages of the introduced possibilistic mixture model-based clustering scheme over conventional approaches. The considered synthetic data were obtained by generating two realizations of 350 samples each. In each realization, the first 100 samples were drawn from the bivariate Gaussian distribution \(f_A\) with

\[
f_A = \mathcal{N}\left(\begin{bmatrix} 0 \\ 3 \\ 2 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}\right)
\]

the second 100 samples were drawn from the bivariate Gaussian distribution \(f_B\) with

\[
f_B = \mathcal{N}\left(\begin{bmatrix} 3 \\ 0 \\ 1 \\ 0.1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}\right)
\]

the third 100 samples were drawn from the bivariate Gaussian distribution \(f_C\) with

\[
f_C = \mathcal{N}\left(\begin{bmatrix} -3 \\ 0 \\ 2 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}\right)
\]

while the final 50 samples were drawn from a bivariate uniform distribution, with each of its components in the interval \([-10, 10]\) (outliers). This way, we obtain a challenging data set, containing a significant proportion of outliers, hence allowing for the objective assessment of the advantages of the introduced possibilistic regard toward generative mixture models.

We postulate a MoG and a PMoG model comprising three component densities (clusters), with full covariance matrices, and we use them to cluster the considered data. In Table 1, we provide the obtained clustering error rates for each density (cluster) (%).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>(f_A)</th>
<th>(f_B)</th>
<th>(f_C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMoG</td>
<td>4.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MoG</td>
<td>7.0</td>
<td>0</td>
<td>48.0</td>
</tr>
<tr>
<td>PFCM</td>
<td>24.0</td>
<td>3.0</td>
<td>52.0</td>
</tr>
</tbody>
</table>

![Fig. 1. Experiment on simulated data: (a) MoG; (b) PMoG; (c) PFCM.](image-url)
error rates. Finally, in Fig. 1, we provide a graphical illustration of the obtained clustering results. We notice that the PMoG algorithm yields considerably different clusters compared to MoG, with their centers being much closer to the actual ones compared to what the competition obtained.

6.2. Topic-based supervised clustering of Flickr images

In our second experiment, we consider an application of the PMM in topic-based image clustering. This application is very interesting since one image might be simultaneously related with multiple topics, and, hence, a clustering scheme inherently allowing for the association of each processed image with multiple topics, and, hence, a clustering scheme inherently allowing for the multiple topic image classifications, we come up with the following convention: we define a threshold $\text{thres}$, and we let each image to be attributed to all the topics yielding a posterior value $q(z_{nk} = 1 | \Psi)$ above that threshold. In Fig. 2, we provide the false positives rate vs. false negatives rate diagram for both the MoG and PMoG models. These plots are obtained by merely repeating our experiments for various values of the decision threshold $\text{thres}$. As we observe, the PMoG model, explicitly designed to handle applications requiring multiple cluster data assignments, such as the examined experimental case, consistently outperforms the MoG model for any performance level (false positives or false negatives rate). We also note that the optimal performance of the PMoG model (optimal combination of false positive and false negative rates) is 32.14%, whereas for the MoG model is 36.54%. For comparison, note that the classification error rate of a linear support vector machine (SVM) classifier [29] in the same problem is 35.93%, that is slightly better than the MoG model.

6.3. UCI data clustering

Finally, we consider an application of the proposed possibilistic clustering approach to generative mixture models in clustering of data sets from the UCI Machine Learning Repository [30]; particularly, we consider the data sets 2Moons, Pima, Magic, and Wine. We fit one MoG and one PMoG model to each one of these data sets, with the number of component densities of the fitted models set equal to the number of clusters we want to obtain, that is the number of classes in each data set case. We consider models with full covariance matrices.

In Table 2, we provide the obtained average error rates of the evaluate models for the considered data sets. As we observe, the proposed approach outperforms the competition, yielding an improved clustering performance for the same computational demands, as the primary computational costs of the EM algorithm for both the MoG and the PMoG model arise from the computation and inversion of the (same) empirical covariance matrices $\Sigma_k$. Additionally, we also compare our method with the possibilistic fuzzy $c$-means algorithm of [14]. As we observe, the EM-based algorithms perform considerably better than PFCM, since their statistical assumptions allow for
the derivation of clusters with much more complex shapes compared to the Euclidean distance-based PFCM.

7. Conclusions

In this work, we introduced a possibilistic approach toward generative mixture models. Our novel approach gives rise to a Binomial allocation of the modeled data over the mixture component densities, and, hence, the derived clusters, that ameliorate significantly the counterintuitive effects induced by the multinomial data allocation assumed by conventional MM formulations, related with the incapacity of such models to distinguish between equal evidence or ignorance, i.e. detecting cases of data points which would be equally likely to belong to more than one clusters, or of data points that would be unlikely to belong to anyone of the known clusters, usually referred to as outliers. This way, the proposed model allows for the effective modeling of data sets where each data point might simultaneously belong to more than one classes, such as in cases of topic-based image retrieval applications, which common MM formulations are not explicitly designed to handle. Experimental evaluation of the PMM was conducted considering the case of a Gaussian clusters selection (PMoG model). We examined a number of applications, including clustering of synthetic data significantly contaminated by outliers, a topic-based image clustering application, where each image might be associated with multiple topics, and a benchmark data clustering application, using data from the UCI repository. In all cases, the proposed approach managed to outperform its natural competitor, i.e. the MoG model, in terms of clustering performance, while imposing the same computational burden. Our future research directives include examining the applicability of the possibilistically-induced Binomial data allocation PMM found in the context of more complex generative models, such as Markov chain and Markov random field models.

References


Sotirios P. Chatzis received the MEng in Electrical and Computer Engineering with distinction from the National Technical University of Athens, in 2005, and the PhD in Machine Learning, in 2008, from the same institution. From January 2009 till June 2010 he was a Postdoctoral Fellow with the University of Miami, USA. He currently is a postdoctoral researcher at the Department of Electrical and Electronic Engineering, Imperial College London. His major research interests comprise machine learning theory and methodologies with a special focus on hierarchical Bayesian models, reservoir computing, robot learning, demonstration capture, corpus, quantum statistics, and Bayesian non-parametrics. His PhD research was supported by the Boddosaki Foundation, Greece, and the Greek Ministry for Economic Development, whereas he was awarded the Dean’s scholarship for PhD studies, being the best performing PhD student of his class. In his first 5 years as a researcher he has first-authored 23 papers in the most prestigious journals of his research field.

Gavrill Tsigchenakis received the PhD degree in electrical and computer engineering in 2003 from the National Technical University of Athens, Greece. From 2004 to 2006, he was a research faculty at the Center for Computational Biomedicine, Imaging and Modeling (CBIM), Rutgers University, Piscataway, NJ, under the supervision of D. Metaxas. During 2007–2008, he was a visiting Assistant Professor at the Electrical and Computer Engineering Department, University of Miami, Miami, FL. From August 2008 until August 2010, he was a senior researcher at the Center for Computational Science, University of Miami. Currently, he is a faculty member of the Department of Computer & Information Science, Indiana University—Purdue University Indianapolis, United States. His research focus is on computer vision and machine learning, with applications to image and video analysis, computational biomechanics, and ocean engineering. His work can be found in journals and peer review conference proceedings, and he serves on the program committees of several conferences and workshops.