Probabilistic Clustering Based on Langevin Mixture

Ola Amayri  
Electrical and Computer Engineering Department  
Concordia University  
Montreal, Canada, H3G 2W1  
Email: o_amayri@ece.concordia.ca

Nizar Bouguila  
Concordia Institute for Information Systems Engineering  
Concordia University  
Montreal, Canada, H3G 2W1  
Email: bouguila@ciise.concordia.ca

Abstract—In this paper, we propose a statistical framework for clustering spherical data which are usually found in machine learning, data mining and computer vision applications. Our framework is based on finite Langevin mixture models which provide a very natural representation of normalized vectors in high dimensional spaces in which the data lie on unit hypersphere. Moreover, we developed minimum message length (MML) criterion for the selection of finite Langevin mixture components from which different probabilistic information divergence distances are then derived. Through empirical experiments, we demonstrate the merits of the proposed learning framework through challenging applications involving spam filtering using visual email content and email categorization.

I. INTRODUCTION

Clustering permits grouping of data into homogenous categories and allows to investigate, model and analyze very complex data and phenomena. The key challenge is to be able to cluster in high dimensional space in spite of redundant, incomplete, sparse and erroneous data. To date, many approaches have been proposed to group and manage generated information and can be broadly grouped into: model-based approaches and similarity approaches. Since clustering is perceived as an unsupervised classification, in this paper we will consider model-based approaches that are capable of revealing enriched presentation for specific desirable characteristic of sought patterns and exploit domain knowledge for clustering. Finite mixture, for instance, is a common model-based approach for data modeling which provides principled and effective way to formally model uncertainty based on unsupervised groupings that disposes similar data to same clusters. Therein, the assignment of points to clusters is defined by the membership of a data point \( X_i \) in a cluster \( j \) and is given as a posterior probability denoted by \( p(j|X_i) \). The underlying probability model and its parameters are estimated using Expectation-Maximization (EM) algorithm in which the determination of the memberships of the data points is tackled. Moreover, cosine similarity between clusters has been shown to provide promising results and better clustering when the data at hand is large and sparse in high-dimensional spaces [1]. Indeed, the cosine of angles between \( L_2 \) normalized feature vectors in high dimensional spaces, which they lie on a unit hypersphere, can be naturally modeled using Langevin models [1]. From application perspective, the adoption of \( L_2 \) normalization has been shown to play an important role, as a preprocessing step, in many practical application generally and in text and image clustering particularly. Indeed, in text clustering, experiments have shown that the bias to the size of the document has been alleviated using \( L_2 \) normalization. In image clustering, \( L_2 \) normalization has proved to increase the robustness to various changes, such as illumination changes [2]. Nonetheless, in some multimedia applications a given object can be modeled as a sequence (bag) of low dimensional vectors, rather than one high-dimensional vector. In this case, each cluster can be represented itself by a mixture model and hence, traditional clustering approaches such as \( K \)-means and its variants can not be deployed. To tackle these problems, in this paper we derive probabilistic distances between Langevin mixtures based on information divergence namely, Kullback-Leibler, Rényi, and Jensen-Shannon which has Bhattacharyya as special case. Moreover, we also propose an MML criterion for model selection when considering Langevin mixtures. The rest of this paper is organized as follows. In Section II we briefly introduce the Langevin mixture. In Section III we propose a complete learning algorithm by developing a MML criterion for model selection and we derive probabilistic distances based on Langevin mixture model. We demonstrate the capability and merits of the proposed approach in Section IV. Finally, we conclude the paper in Section V.

II. FINITE LANGEVIN MIXTURE

Let \( \tilde{X} = (X_1, \ldots, X_p) \) be a random unit vector in \( \mathbb{R}^p \). \( \tilde{X} \) is said to have a \( p \)-variate Langevin distribution if its probability density function is given by [3]:

\[
p_p(\tilde{X}|\tilde{\mu}, \kappa) = \frac{\kappa^\frac{p}{2} - 1}{(2\pi)^\frac{p}{2} I_{\frac{p}{2} - 1}(\kappa)} \exp\{\kappa \tilde{\mu}^T \tilde{X}\}
\]  (1)

on the \((p - 1)\)-dimensional unit sphere \( S^{p-1} = \{\tilde{X} | \tilde{X} \in \mathbb{R}^p : ||\tilde{X}|| = (\tilde{X}^T \tilde{X}) = 1\} \), with mean direction unit vector \( \tilde{\mu} \in S^{p-1} \), where \( \tilde{\mu}^T \) denotes the transpose of \( \tilde{\mu} \) and non-negative real concentration parameter \( \kappa \geq 0 \). Furthermore, \( I_p(\kappa) \) denotes the modified Bessel function of first kind and order \( p \) [3]. From Eq. 1 we can notice that Langevin distribution is a member of (curved)-exponential family of order \( p \), whose shape is symmetric and unimodal, with minimal canonical parameter \( \kappa \tilde{\mu} \) and minimal canonical statistic \( \tilde{X} \). Let \( p(\tilde{X}_j|\theta) \) be a mixture of \( M \) Langevin distributions:

\[
p(\tilde{X}_j | \theta) = \sum_{j=1}^{M} p_p(\tilde{X}_j | \theta_j)p_j
\]  (2)
where $\Theta = \{\bar{P} = (p_1, \ldots, p_M), \bar{\theta} = (\theta_1, \ldots, \theta_M)\}$ denotes all the parameters of the mixture model such that $\theta_j = (\mu_j, \kappa_j)$ and $\bar{P}$ represents the vector of clusters probabilities (i.e., mixing weights) which must be positive and sum to one.

III. THE PROPOSED MODEL

A. Model Selection Using MML

One of the central issues in mixture models is to determine the optimal degree of complexity (i.e. optimal number of clusters). A well-known approach that has been shown to be efficient for a variety of mixture models is MML which aims at minimizing the following message length [4]:

$$MessLength(M) \simeq - \log(h(\Theta)) - \log(p(X|\Theta))$$
$$+ \frac{1}{2} \log(|F(\Theta)|) + \frac{N_p}{2}(1 - \log(12))$$

where $h(\Theta)$ is the prior probability, $p(X|\Theta)$ is the likelihood, $F(\Theta)$ is the expected Fisher information matrix, $|F(\Theta)|$ is its determinant, and $N_p$ is the number of free parameters to be estimated which is equal to $M(p+1)-1$ in our case. In the following, we calculate the Fisher information and we propose a prior to obtain the complete message length expression for a finite Langevin mixture.

1) Fisher Information: We replace $F(\Theta)$ by the complete data Fisher information matrix which has a block diagonal structure. Thus, the Fisher information can be approximated as follows:

$$|F(\Theta)| \simeq |F(\bar{P})| \prod_{j=1}^{M} |F(\bar{\mu}_j, \kappa_j)|$$

where $|F(\bar{P})|$ is the determinant of the information matrix of the mixing parameter $\bar{P}$ and we can easily show that [4]:

$$|F(\bar{P})| = \frac{N^{M-1}}{\prod_{j=1}^{M} \mu_j}$$

$|F(\bar{\mu}_j, \kappa_j)|$ is the Fisher information of the Langevin distribution representing component $j$ and we can show that [5]:

$$|F(\bar{\mu}_j, \kappa_j)| = N^{p-2} u^2(\kappa_j) v^2(\bar{\mu}_j, 0)$$

where $n_j$ is the number of vectors assigned to cluster $j$, $u(\kappa_j) = \kappa_j^{(p-2)} A_p(\kappa_j) \frac{(p-1)}{2} \left( \kappa_j - A_p(\kappa_j) \right)$, $v(\bar{\mu}_j, 0) = \prod_{d=1}^{p-1} \sin^{\mu_{d-1}} \mu_{j, d-1}$, where $\bar{\mu}_j, 0 = (\mu_{j,0}, \ldots, \mu_{j,p-1})$ denotes the spherical polar coordinates of $\bar{\mu}_j$ and $A_p(\kappa_j) = t_{\frac{1}{2}p-1}(\kappa_j)$.

2) The Prior $h(\Theta)$: In the absence of any other knowledge about $\bar{P}, \bar{\mu}_j$ and $\kappa_j$, we suppose that they are mutually independent, which yields to the following prior distribution over the parameters:

$$h(\Theta) = h(\bar{P}) \prod_{j=1}^{M} h(\kappa_j)$$

For the mixing probabilities $\bar{P}$, a common choice as a prior is the Dirichlet distribution, which is reduced to uniform distribution considering certain parameters value [4]:

$$h(\bar{P}) = (M-1)!$$

For the parameter $\bar{\mu}_j$, we consider a uniform prior on the surface $S_p$ of the unit $(p-1)$-sphere:

$$h(\bar{\mu}_j) = \frac{1}{S_p}, \quad S_p = \begin{cases} \frac{p}{2}, & \text{if } p \text{ is even} \\ \frac{p+1}{p-1} \frac{p}{2}, & \text{if } p \text{ is odd} \end{cases}$$

Furthermore, we consider the following prior, which has been found appropriate according to our experimental results, for the concentration parameter of a Langevin distribution [6]:

$$h(\kappa_j) = \frac{N^{p-2}}{(1 + \kappa_j^2)^{p+1}}$$

By substituting Eq.7 and Eq.4 into Eq.3, we obtain the MML criterion for a finite mixture of Langevin distributions. The complete algorithm to learn the Langevin mixture is finally summarized below. For each candidate value of $M$ do:

- INITIALIZATION: apply spherical K-means [7] on $N p$-dimensional vectors to obtain the initial parameters for each component $\mu_j$, $\kappa_j$, $j = 1, \ldots, M$
- Estimate the parameters of the Langevin mixture using the algorithm in [1].
- Calculate the associated criterion $MessLength(M)$ using Eq. 3
- Select the optimal model $M^*$ such that $M^* = \arg \min MessLength(M)$

B. Probabilistic Distance Clustering

After Langevin mixtures have been learnt for every category, we need to group similar documents (e.g. images, text, videos) together. To achieve this goal, we need to calculate the similarity (or dissimilarity) distance between different mixtures of Langevin distributions by developing different information divergence measures. In the following, we propose some probabilistic distances based on Langevin distribution$^1$.

1) Kullback-Leibler Divergence (KL): The symmetric KL between two probability distributions $p(X|\Theta)$ and $q(X|\Theta)$ is given by:

$$D_{KL}(p(X|\Theta), q(X|\Theta)) = KL(p(X|\Theta), q(X|\Theta))$$

$$+ KL(q(X|\Theta), p(X|\Theta))$$

where $KL(p(X|\Theta), q(X|\Theta)) = \int q(X|\Theta) \log \frac{p(X|\Theta)}{q(X|\Theta)} dX$. The fact that the Langevin distribution belongs to the exponential family of distributions allows us to find a closed-form expression for the KL divergence between two Langevin

$^1$Detailed derivations of proposed distance have been omitted due to the space limit.
distributions:

\[ D_{KL}(p(\vec{X}|\theta), q(\vec{X}|\theta)) = -\log \frac{\kappa^\frac{\alpha}{2} - 1}{(2\pi)^{\frac{\alpha}{2}} I_{\frac{\alpha}{2} - 1}(\kappa)} + \log \frac{\kappa^\frac{\beta}{2} - 1}{(2\pi)^{\frac{\beta}{2}} I_{\frac{\beta}{2} - 1}(\kappa)} + [\kappa\vec{\mu} - \kappa\vec{\mu}]^T \vec{d}(\kappa)\vec{\mu} \]  

However, a closed form expression does not exist in the case of finite mixture models. Thus, we propose the use of different sampling approaches that have been proposed in [8] (KLMA) and [9] (KLVAR, KLUFP) for GMM.

2) Jensen-Shannon Divergence (JS): JS is an alternative divergence that measures the probability that two samples where drawn from the same distribution. JS between two distributions \( p(\vec{X}|\theta) \) and \( q(\vec{X}|\theta) \) is given by:

\[ D_{JS}(p(\vec{X}|\theta), q(\vec{X}|\theta)) = H[\beta p(\vec{X}|\theta) + (1 - \beta)q(\vec{X}|\theta)] - \beta H[p(\vec{X}|\theta)] - (1 - \beta)H[q(\vec{X}|\theta)] \]  

where \( \beta \) is a parameter and \( H[p(\vec{X}|\theta)] = -\int p(\vec{X}|\theta) \log p(\vec{X}|\theta) d\vec{X} \) is the Shannon entropy of \( p(\vec{X}|\theta) \). It is clear that when \( \beta = \frac{1}{2} \), the JS divergence is average the distance of Kullback-Leibler. Thus \( H[p(\vec{X}|\theta)] \) Shannon entropy for Langevin distribution is given by:

\[ H[p(\vec{X}|\theta)] = d_{\vec{\mu}}(\kappa) + \frac{1}{(2\pi)^{\frac{\alpha}{2}} I_{\frac{\alpha}{2} - 1}(\kappa)} \]  

3) Rényi Divergence (RD): RD is another divergence measure between \( p(\vec{X}|\theta) \) and \( q(\vec{X}|\theta) \):

\[ D_{RD}(p(\vec{X}|\theta), q(\vec{X}|\theta)) = \left[ \int_{\Omega} p(\vec{X}|\theta)^{\sigma} q(\vec{X}|\theta)^{1-\sigma} d\vec{X} \right]^{\frac{1}{\sigma}} \]  

where \( \sigma \) controls the amount of smoothing for the distributions, \( \sigma > 0 \) and \( \sigma \neq 1 \). In case of Langevin distribution, we find closed form of RD of order \( \sigma \)

\[ D_{RD}(p(\vec{X}|\theta), q(\vec{X}|\theta)) = \left[ \frac{\kappa^\frac{2}{2} - 1}{(2\pi)^{\frac{2}{2}} I_{\frac{2}{2} - 1}(\kappa)} \right]^{\sigma} \times \left[ \frac{\kappa^\frac{2}{2} - 1}{(2\pi)^{\frac{2}{2}} I_{\frac{2}{2} - 1}(\kappa)} \right]^{1-\sigma} \left[ \frac{(2\pi)^{\frac{2}{2}} I_{\frac{2}{2} - 1}(\xi_\kappa, \kappa)}{\xi_\kappa, \kappa} \right]^{\frac{2}{2} - 1} \]  

where \( \xi_\kappa, \kappa = \sqrt{\kappa^2 + \kappa^2 + 2\kappa^2 \beta} \) and \( \tau_\mu, \kappa = \frac{\kappa^2 + \kappa^2 \beta}{\xi_\kappa, \kappa} \) are the concentration parameter and mean direction defining the Langevin distribution, respectively, that results from the multiplication of two Langevin distributions: \( p_\mu(X|\mu, \kappa)p_\kappa(X|\kappa) \propto p_\mu(X|\tau_\mu, \kappa, \xi_\mu, \kappa) \). It is worth stressing that when \( \sigma = \frac{1}{2} \), RD is reduced to Bhattacharyya divergence (Bhatt). Because of the absence of closed form expressions for mixture models in case of JS, RD and Bhatt, we used Monte Carlo simulation in our experiments.

IV. EXPERIMENTAL RESULTS

A. Email Categorization

Email categorization is a rich and multifarious problem that poses several challenges. In particular, email folders may vary across different users and more importantly are richer than simple semantic topics since they may correspond to project groups, certain recipients, etc. Indeed, the manner that email users organize their files might change overtime, for instance, users may create new folders, while stop using already existing ones. The goal of this first application is to validate the efficiency of proposed learning algorithm of Langevin finite mixture models (See algorithm in III) and then compare its performance to other generative model that was widely used in the past namely GMM. We conducted our experiments on a challenging dataset that has been widely considered in the past called Enron Email dataset and is composed of 200,399 emails belonging to 158 users. We conducted experiments on the largest email directories that have been used in the past. Each of those directories has subfolders, where we characterize them as topical and non-topical (i.e. computer generated). We removed the non-topical (i.e. computer generated) folders from all the directories. Next, we flatten all the folder hierarchies and removing all folders that contains less than three emails. Then, we start presenting the dataset in vector space. First, we start by tokenizing emails and then we build dataset dictionary where all stop words and words which occur less than three times are removed. Second, we present each document by a vector of counts, which is then \( \bar{L}_2 \) normalized, and start clustering. Both Langevin mixtures and GMM were able to find the exact number of clusters. This can be explained by the fact that MML criterion uses prior information that allows better comprehension to the application and data at hand. Table I shows the average confusion matrices for clustering using Langevin mixture and GMM, respectively. According to the results, when using Langevin mixture the average clustering accuracy was 93.29% ± 1.4, which actually better than the accuracy achieved by GMM 84.78% ± 0.9.

B. Spam Filtering

Email spam is a major problem in electronic communications. Particularly, researchers have figured out that text-based techniques might be ineffective because of a novel spammers trick namely image-based spam (i.e. email which includes embedded image). Thus, some approaches have been proposed to detect the nature of email from its image content. The majority of the approaches consider, however, only the textual content of the image and ignore its rich low-level visual content (e.g. color, texture, shape) which can be very helpful as clearly shown for instance in previous works about content-based image indexing and retrieval. Recently, authors in [10] proposed clustering of image, based on GMMs using the agglomerative information bottleneck principle for initialization, where each image is represented as a set of feature vectors (one vector for each image’s pixel) and then using the
Jensen-Shannon divergence between GMMs to classify new emails. The main goal of this experiment is to investigate the probabilistic distances between Langevin mixtures, which we proposed in section III-B, by following the approach proposed in [10]. For fair comparison, experiments were conducted on datasets that were used recently in [10], namely, Princeton dataset. Princeton dataset has 1071 emails which spread into 178 categories. Following [10], we start by resizing all images to 100 × 100 pixels, which has been shown to have reasonable computation time and it is robust to pixels scaling and randomization. Then, we start extracting visual features from each image. In particular, we presented each pixel in the image by a vector of seven tuples: two parameters for pixel coordinates (x, y), three for (L*, a*, b*) color attributes and two for texture attributes for anisotropy and contrast features. Then, we L2 normalize each vector, as a result, we can consider each pixel in a mixture of Langevin distributions using algorithm in Section III. To evaluate the performance of spam clustering framework, we calculate the clustering accuracy (Acc = TP + TN / TP + FP + TN + FP). A summary of the clustering as displayed in Table II clearly shows that our framework based on Langevin mixture model generally achieves superior performance than that in [10]. Note that, even for JS, which was adopted in [10] Langevin mixture outperform JS based on GMM where the best result achieved was 84%. This can be justified by the fact that the GMM, which clustering is based implicitly on the Euclidean distance or Mahalanobis, is inadequate for characterizing L2 normalized data which clustering structure is better uncovered by considering the cosine similarity as assumed by the Langevin mixture especially when dealing with the problem of spam filtering [11].

TABLE I

<table>
<thead>
<tr>
<th>Probabilistic distance</th>
<th>Langevin Mixture</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLVAR</td>
<td>90.55</td>
<td>82.01</td>
</tr>
<tr>
<td>KL_DPP</td>
<td>83.45</td>
<td>75.84</td>
</tr>
<tr>
<td>KL_SMA</td>
<td>81.79</td>
<td>74.19</td>
</tr>
<tr>
<td>Bhattacharyya</td>
<td>89.91</td>
<td>82.90</td>
</tr>
<tr>
<td>RD</td>
<td>90.82</td>
<td>76.91</td>
</tr>
<tr>
<td>JS</td>
<td>91.01</td>
<td>82.99</td>
</tr>
</tbody>
</table>

TABLE II

SPAM CLASSIFICATION ACCURACY (IN %) FOR DIFFERENT PROBABILISTIC DISTANCE BASED ON LANGEVIN MIXTURE AND GMM.

V. CONCLUSION

In this paper, we have developed an MML criterion to tackle the challenging problem of selecting the number of components in an unsupervised fashion. Moreover, several probabilistic measures have been proposed to tackle the problem of objects classification when each object is represented by a bag of vectors modeled via finite Langevin mixture. Empirical experiments have shown that proposed framework achieves promising results and outperforms the extensively used GMM on challenging real problems. A potential future work can be devoted to developing hybrid generative discriminative models via the generation of SVM kernels from Langevin mixture.

ACKNOWLEDGEMENTS

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

REFERENCES


3Available at http://www.princeton.edu/cass/spam/spam_bench/