Model-based approach for high-dimensional non-Gaussian visual data clustering and feature weighting

Tarek Elguebaly [a], Nizar Bouguila [b],*

[a] *Electrical and Computer Engineering (ECE), Concordia University, Montreal, QC H3G 2W1, Canada*
[b] *Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 2W1, Canada*

A B S T R A C T

Clustering is the task of classifying patterns or observations into clusters or groups. Generally, clustering in high-dimensional feature spaces has a lot of complications such as: the unidentified or unknown data shape which is typically non-Gaussian and follows different distributions; the unknown number of clusters in the case of unsupervised learning; and the existence of noisy, redundant, or uninformative features which normally compromise modeling capabilities and speed. Therefore, high-dimensional data clustering has been a subject of extensive research in data mining, pattern recognition, image processing, computer vision, and other areas for several decades. However, most of existing researches tackle one or two problems at a time which is unrealistic because all problems are connected and should be tackled simultaneously. Thus, in this paper, we propose two novel inference frameworks for unsupervised non-Gaussian feature selection, in the context of finite asymmetric generalized Gaussian (AGG) mixture-based clustering. The choice of the AGG distribution is mainly due to its ability not only to approximate a large class of statistical distributions (e.g. impulsive, Laplacian, Gaussian and uniform distributions) but also to include the asymmetry. In addition, the two frameworks simultaneously perform model parameters estimation as well as model complexity (i.e., both model and feature selection) determination in the same step. This was done by incorporating a minimum message length (MML) penalty in the model learning step and by fading out the redundant densities in the mixture using the rival penalized EM (RPEM) algorithm, for first and second frameworks, respectively. Furthermore, for both algorithms, we tackle the problem of noisy and uninformative features by determining a set of relevant features for each data cluster. The efficiencies of the proposed algorithms are validated by applying them to real challenging problems namely action and facial expression recognition.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Clustering, a common technique for statistical data analysis, is one of the most important tools to find a structure in a collection of unlabeled data. It is used to group items that seem to fall naturally together [1]. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Lately, the problem of clustering data into homogeneous groups has been widely studied due to its significance in a variety of areas such as image processing, data mining, and computer vision. The vast majority of these approaches are distance-based algorithms which partition the data set into subsets according to some defined distance measure. However,

these algorithms perform poorly on high-dimensional data because distances between points become more uniform [2]. Therefore, some subspace clustering approaches were introduced in order to address this concern. Though, these approaches make strong assumptions about the distributions of the data, and rely on heuristics to find clusters satisfying the assumptions [3]. Furthermore, in high-dimensional data sets, traditional clustering algorithms tend to break down both in terms of accuracy, as well as efficiency, so-called curse of dimensionality [4]. Here, we present an alternative algorithm based on finite mixture model designed especially for high-dimensional clustering that makes reasonable distributional assumptions and provides satisfying theoretical guarantees.

Finite mixture models provide a natural representation of heterogeneity in a finite number of latent classes by modeling a statistical unknown distribution by a weighted sum of other distributions. Thus, lately, finite mixture models have been widely used to provide a formal framework for clustering. However, there are several challenges that should be handled when using finite mix-

ture models such as the choice of the statistical distribution that represents each group or cluster and the learning algorithm used for mixture's parameters estimation, the selection of the model order (i.e. number of clusters), and the number of relevant features in high dimensional data. The main objective of this paper is to sum up all these challenging interrelated problems in one unified model. In the field of finite mixtures, the mixture of Gaussians is generally used for its simplicity, nonetheless, it has been observed that the Gaussian distribution is unsuitable for modeling data in complex real life applications [5] and especially in the case of computer vision problems [6]. Therefore, in order to overcome the rigidity of the Gaussian distribution, researchers were encouraged to exploit various distributions such as the generalized Gaussian distribution (GGD). The GGD is a continuous probability distribution capable of modeling data with different shapes [7–9]. The GGD includes the Laplacian and the Gaussian as special cases as well as the uniform distribution as limiting case [10] and has been employed in many challenging problems (see, for instance [11–14]). Even with the higher flexibility that GGD offers, it is still a symmetric distribution inappropriate to model non-symmetrical data. In this article, we suggest the use of the asymmetric generalized Gaussian distribution (AGGD) capable of modeling non-Gaussian asymmetrical data [15]. The AGGD has two variance parameters for left and right parts of the distribution, which allow it not only to approximate a large class of statistical distributions (e.g. impulsive, Laplacian, Gaussian and uniform distributions) but also to include the asymmetry.

A standard method to learn finite mixture models is maximum likelihood which generally estimates the parameters through the expectation maximization (EM) framework. The EM algorithm enables us to update the mixture parameters with respect to a data set. In order to use the EM algorithm, an appropriate number of clusters should be predefined, otherwise, the EM algorithm will lead to a poor result. Thus, another important part of the mixture modeling problem concerns determining the number of consistent components which best describes the data. For this purpose, many approaches have been suggested for automatic selection of the number of clusters (see, for instance, [16–18]). Examples of selection criteria include Akaike information criterion (AIC) [19], minimum description length (MDL) [20] and Laplace empirical criterion (LEC) [21]. These criteria, however, may overestimate or underestimate the number of clusters. Furthermore, they are time-consuming and may lead to a sub-optimal solution because model selection and parameters estimations are determined in two separate steps. In our first algorithm, we propose a learning method that can integrate simultaneously parameter estimation and model selection by implementing MML criterion [22] into an EM algorithm [23].

In [24], the authors used another clustering method, Rival Penalized Competitive Learning (RPCL), which could detect the real number of clusters with little prior knowledge. In RPCL, the assigned number of clusters must be bigger than the true number of clusters, then, the algorithm can automatically select the correct cluster number by gradually driving extra seed points far away from the input data set. However, its performance is sensitive to the selection of the de-learning rate, such that if it is not well selected, the RPCL may completely break down. In order to overcome this problem, the Rival Penalized EM (RPEM) algorithm was proposed for density mixture clustering [25]. The RPEM learns the model parameters by making the mixture components compete with each other at each time step; this can be done by not only updating the winning density component parameters to adapt to the input but also all rivals parameters are penalized with the strength proportional to the corresponding posterior density probabilities. Therefore, the RPEM is able to automatically select an appropriate number of densities by fading out the redundant

densities from a density mixture which can save the computing time. To the best of our knowledge, this is the first attempt to employ the RPEM algorithm for AGGM model selection and parameter learning.

Feature selection, the task of identifying relevant or discriminative features, represents an essential step in high-dimensional data clustering. This is actually an important step, since the main goal is not only the determination of clusters and their parameters but also to provide the most parsimonious model that can accurately describe the data [26,27]. Furthermore, feature selection can speed up learning and improve model accuracy and generalization. Therefore, the selection of relevant features in multidimensional data represents a major concern in several image processing, computer vision and pattern recognition applications such as object detection [28], handwriting separation [29], image retrieval, categorization and recognition [30]. However, the majority of research in mixture models assume that all features have the same weight and use a pre-processing step such as principal components analysis to transform the original features into a new dimension-reduced space. The main drawback of that approach is that the physical meaning of the original features is generally lost [31]. Moreover, the learning of the mixture parameters (i.e. both model selection and parameters estimation) is greatly affected by the quality of the features used as shown for instance in [32,33]) where the authors considered the Gaussian assumption by assuming diagonal covariance matrices for all clusters. The work of [34] proposed a new feature selection method using Markov Blanket capable of eliminating irrelevant and redundant features. They integrated feature selection with RPEM for Gaussian mixtures clustering. Therefore, data clustering, model learning and feature selection were performed in a single learning algorithm. In [35], a novel approach to combine clustering and feature selection by implementing a wrapper strategy for feature selection was presented. Thus, features were directly selected by optimizing the discriminative power of the used partitioning algorithm. The authors in [36] presented an algebraic approach to variable weighting by maximizing a score based on the spectral properties of the kernel matrix. This work presented a definition of relevancy based on spectral properties of the Laplacian of the features measurement matrix. Then, the feature selection process is based on a continuous ranking of the features defined by a least-squares optimization process. Zeng et al. [37] associated a weight to each feature or kernel and incorporated it into the built-in regularization of the Learning-Based Clustering (LLC) [38] algorithm to take into account the relevance of each feature or kernel for clustering. Their idea was that the resulting weighted regularization with an additional constraint on the weights is equivalent to a known sparse-promoting penalty. Hence, the weights of those irrelevant features or kernels can be shrunk toward zero. In this article, and following recent approaches (see, for instance [30,32]), we approach the feature selection problem in unsupervised learning by casting it as an estimation problem, thus avoiding any combinatorial search. For each feature, we associate a relevance weight which measures the degree of its dependence on class labels.

The remainder of the paper is structured as follows: In Section 2 we review the feature selection model based on the asymmetric generalized Gaussian mixture. In Section 3, we propose our first learning algorithm by implementing MML criterion into an EM algorithm. In Section 4, we integrate the concept of feature saliency into the RPEM algorithm for the AGGM model. In Section 5, we assess the performance of our approach for action and facial expression recognition; while comparing them to other state-of-the-art methods. Our last section is devoted to the conclusion and some perspectives.