Pattern Recognition Letters 32 (2011) 1154–1161

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An entropy weighting mixture model for subspace clustering of high-dimensional data

Liuqing Peng *, Junying Zhang

School of Computer Science and Technology, Xidian University, 2, Taibai Road, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 17 July 2010
Available online 12 March 2011
Communicated by L. Heutte

Keywords:
Subspace clustering
High-dimensional data
Gaussian mixture models
Local feature relevance
Shape volume

ABSTRACT

In high-dimensional data, clusters of objects usually exist in subspaces; besides, different clusters probably have different shape volumes. Most existing methods for high-dimensional data clustering, however, only consider the former factor. They ignore the latter factor by assuming the same shape volume value for different clusters. In this paper we propose a new Gaussian mixture model (GMM) type algorithm for discovering clusters with various shape volumes in subspaces. We extend the GMM clustering method to calculate a local weight vector as well as a local variance within each cluster, and use the weight and variance values to capture main properties that discriminate different clusters, including subsets of relevant dimensions and shape volumes. This is achieved by introducing negative entropy of weight vectors, along with adaptively-chosen coefficients, into the objective function of the extended GMM. Experimental results on both synthetic and real datasets show that the proposed algorithm outperforms its competitors, especially when applying to high-dimensional datasets.

1. Introduction

Clustering on high-dimensional data has encountered great challenges. As we know, clustering analysis seeks to find groups or clusters of similar objects based on a distance/similarity measure. However, for high dimensional datasets, such as in the applications of image processing, microarray analysis, and text clustering, not all dimensions/features are important to discover a cluster and many of the dimensions are often irrelevant (Parsons et al., 2004). Further, due to the specificity of a cluster, different clusters usually have different subsets of important dimensions. All of these facts cause traditional distance/similarity measures that use all dimensions with equal relevance become ineffective (Domeniconi et al., 2007).

A few recently developed approaches referred to as *subspace clustering*, have been put forward for high-dimensional data clustering. In the approaches, clusters are found in subspaces rather than in the entire data space. And different clusters are allowed to exist in different subspaces. According to the ways that the subspaces of clusters are determined, subspace clustering approaches are further divided into two types, hard subspace clustering and soft subspace clustering (Jing et al., 2007).

Hard subspace clustering methods (Agrawal et al., 1998, 1999, 2005; Cheng et al., 1999; Kailing et al., 2004; Parsons et al., 2004) use heuristic criterions to find subspaces of clusters. Generally, they search a few relevant dimensions within each cluster according to some heuristic criterions. The subspace of a cluster thus is the direct combination of the searched dimensions. Main disadvantages of the methods include: certain key parameters in the heuristic criterions are difficult for users to master and set; moreover, clustering results are sensitive to changes in these parameters (Jing et al., 2007; Parsons et al., 2004).

Soft subspace clustering methods (Domeniconi et al., 2004, 2007; Friedman and Meulman, 2004; Frigui and Nasraoui, 2004; Jing et al., 2005, 2007) determine subspaces of clusters according to the relevance of dimensions in discovering the corresponding clusters. They associate to each cluster a local weight vector, with each weight value capturing the contribution degree of a dimension in identifying the cluster. Large weight values are obtained for relevant dimensions along which objects are tightly distributed. Accordingly, the subspace of a cluster is the weighted combination of dimensions by the corresponding weight vector. By using different weight vectors for different clusters locally, a soft subspace clustering gives a detailed description for a dataset.

In addition to subspace clustering, recently a few feature-selection oriented clustering methods are proposed for high-dimensional data clustering, where a common global feature weight vector is used to discriminate clusters. In (Tsai and Chiu, 2008), a feature weight self-adjustment (FWSA) mechanism is proposed to assign features with different weight values, according to their ability in identifying the same clusters and distinguishing different clusters.

* Corresponding author. Tel.: +86 13468817501; fax: +86 29 88203692.
E-mail address: pengann@163.com (L. Peng).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.03.003

In (Law et al., 2004), a minimum message length (MML) criterion is embedded into mixture-based clustering model, to simultaneously estimate the feature saliencies and the number of clusters. Our recent study, however, shows that the MML clustering method may become ineffective if some degree of overlap exists between any two clusters. This is probably caused by the disability of the MML criterion under certain circumstances.

In (Li et al., 2009), the concept of local feature saliency is developed. Unlike the previous two methods, the new one obtains different feature vectors for different clusters; thus is much more like a subspace clustering method. The method also estimates the number of clusters. However, the same result of wrong cluster number being produced would occur when clusters overlap.

For most soft subspace clustering methods such as Domeniconi et al. (2004, 2007) and Jing et al. (2005, 2007), properties of *shape volumes*, however, are not considered or distinguished among clusters, which usually causes the wrong assignments of boundary objects. Fig. 1 gives a simple example. Fig. 1a depicts the original two clusters of objects elongated along the x and y dimensions. Fig. 1b shows the transferred clusters, with the relative position of an object to its cluster center transferred by the corresponding weight vector. It can be seen that the weight vectors reflect the relevance of dimensions and reshape each cluster as a spherical cloud (Domeniconi et al., 2007). However, the transferred clusters are different in their shape volumes. This causes a simple distance measure, like Euclidean distance, is not capable of partitioning the dataset well, especially partitioning those boundary objects.

The problem can be solved by applying Gaussian mixture models (GMMs). The main reason for the above problem is that some simple models like K-means are used. In the models, there is no parameter available for capturing the shape volume property of a cluster. Hopefully in GMMs, each covariance matrix parameter describes the shape property of a cluster, which surely covers the shape volume property. One can change the form of covariance matrix of a cluster, to obtain a local weight vector as well as a local variance, and use the second local variance parameter to capture the shape volume property of the cluster. Fig. 1c shows the new sample distribution after cluster transformation by both the local weight vectors and the local variances of different clusters. It can be seen that the transferred clusters are much more comparable with each other in their shape volumes, which makes the partition of the objects become much easier.

So in this paper, we propose a new GMM type soft subspace clustering algorithm. By constructing a new form of covariance matrix with a local weight vector and a local variance in each Gaussian component of a GMM, we extend GMM clustering method and obtain the corresponding objective function. For estimating parameters in the new model, we apply the scheme of introducing negative entropy of weight vectors into the objective function. In addition, we use different coefficients when different negative entropy items are added into the objective function, and an adaptive way is presented to obtain these coefficients. We evaluate the efficiency of the new mixture model type algorithm using synthetic and real datasets, and compare its classification performance with the performance of other mixture model type algorithms.

The remainder of this paper is organized as follows: Section 2 begins with an overview of GMMs; further the new special model used in our algorithm is presented. In Section 3, we provide an effective subspace clustering algorithm by introducing negative entropy of weight vectors. In Section 4, the synthetic and real data experimental evaluation of our model is conducted. The final section summarizes the results of the paper.

2. Model formulation

2.1. Gaussian mixture models

As a fundamental model hypothesis, Gaussian mixture is typical and well studied (Dasgupta, 1999). In the model, data objects are thought of as originating from various sources or components, and each source is modeled by a Gaussian distribution. Suppose that we have a set of objects $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ $(\boldsymbol{x}_i \in \boldsymbol{R}^d)$ that were drawn from a GMM comprised of K Gaussian components, the probability density function (pdf) on point $\boldsymbol{x}_i$ is given by

$$\Phi(\boldsymbol{x}_i; \boldsymbol{\Theta}_0) = \sum_{k=1}^{K} \alpha_k \phi(\boldsymbol{x}_i | \boldsymbol{c}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where α_k is the mixing weight of component k, with $\sum_{k=1}^{K} \alpha_k^d = 1$ and $0 \leqslant \alpha_k \leqslant 1$. $\boldsymbol{c}_k$ is the mean of objects in component k, and $\boldsymbol{\Sigma}_k$ is the covariance matrix. $\boldsymbol{\Theta}_0$ is defined as $\{(\alpha_k, \boldsymbol{c}_k, \boldsymbol{\Sigma}_k) | 1 \leqslant k \leqslant K\}$. $\phi(\boldsymbol{x}_i | \boldsymbol{c}_k, \boldsymbol{\Sigma}_k)$ is the pdf of the kth Gaussian component, and can be further expressed as

$$\phi(\boldsymbol{x}_i | \boldsymbol{c}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{c}_k - \boldsymbol{x}_i)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{c}_k - \boldsymbol{x}_i)\right). \tag{2}$$

Parameters of the model can be estimated by EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997).

2.2. Subspace clustering model

For a high-dimensional data clustering task, we extend the GMM by constructing a new special form of covariance matrix for each component (or cluster) as

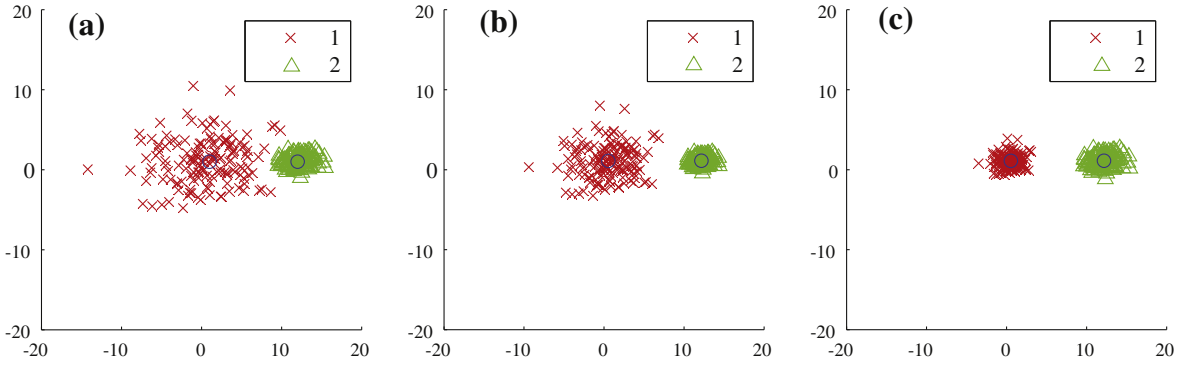

Fig. 1. Illustration of the roles of local weight vectors and local variances in soft subspace clustering. (a) Clusters in original input space. (b) Clusters transformed by the local weight vectors. (c) Clusters transformed by both the local weight vectors and the local variances.