Pattern Recognition 45 (2012) 1146–1158

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Fuzzy Gaussian Mixture Models

Zhaojie Ju *, Honghai Liu

Intelligent Systems and Biomedical Robotics Group, School of Creative Technologies, University of Portsmouth, PO12DJ, UK

ARTICLE INFO

Article history:
Received 14 December 2010
Received in revised form
26 August 2011
Accepted 28 August 2011
Available online 19 September 2011

Keywords:
Fuzzy C-means
Gaussian Mixture Models
Fuzzy Gaussian Mixture Models
EM algorithm

ABSTRACT

In this paper, in order to improve both the performance and the efficiency of the conventional Gaussian Mixture Models (GMMs), generalized GMMs are firstly introduced by integrating the conventional GMMs and the active curve axis GMMs for fitting non-linear datasets, and then two types of Fuzzy Gaussian Mixture Models (FGMMs) with a faster convergence process are proposed based on the generalized GMMs, inspired from the mechanism of Fuzzy C-means (FCMs) which introduces the degree of fuzziness on the dissimilarity function based on distances. One is named as probability based FGMMs defining the dissimilarity as the multiplicative inverse of probability density function, and the other is distance based FGMMs which define the dissimilarity function focusing the degree of fuzziness only on the distances between points and component centres. Different from FCMs, both of the proposed dissimilarity functions are based on the exponential function of the distance. The FGMMs are compared with the conventional GMMs and the generalized GMMs in terms of the fitting degree and convergence speed. The experimental results show that the proposed FGMMs not only possess the non-linearity to fit datasets with curve manifolds but also have a much faster convergence process saving more than half computational cost than GMMs'.

1. Introduction

As one of the most statistically mature methods for clustering [1–5], Gaussian Mixture Models (GMMs) are used intensively in object tracking [6,7], background subtraction [8,9], feature selection [10,11], signal analysis [12,13] and learning and modelling [14–18]. Various kinds of GMMs based methods are developed for specific applications, such as adapted GMMs [19], which are used for dealing with undesired effects of variations in speech characteristics, Mahalanobis distance based GMMs [20], which are capable of splitting one component into two new components, and Wrapped Gaussian Mixture Models [21] using an expectation–maximization algorithm suitable for circular vector data to model dispersion phases. However, more components are required when fitting the datasets with non-linear manifolds because of the intrinsic linearity of Gaussian model which leads to relative large fitting error. To solve this problem and approximate datasets with curve manifolds better, Zhang et al. [22] proposed active curve axis Gaussian Mixture Models (AcaGMMs) which are non-linear probability models. PCA and least-squares fitting methods are used to 'bend' AcaGMMs in the principal plane and points are considered by handling the projection points on the principal axis.

Fuzzy C-means (FCMs), also known as fuzzy ISODATA, was developed by Dunn in 1973 [23] and improved by Bezdek in 1981 [24]. It is a popular and effective clustering method which employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. It employs a weighting exponent m on each fuzzy membership and distances between points and centres. The effects of the weighting exponent are discussed that optimal m may result in better performance or fast convergence in [25,26], and approaches of determining the weighting exponent have also been presented in [25,27,28]. The algorithmic frameworks of FCM and GMMs are closely related [29,30]. Based on FCM, Gustafson et al. [31] defined fuzzy covariance matrices of clusters which means that different clusters in the same dataset may have different geometric shapes. To define these different geometric shapes of clusters, Tran et al. [32] further made a modification of GMMs for speaker recognition, which refined the distances in the FCM functions as the negative of logarithms of density functions. Therefore, the relationship between the membership and distance is transferred from exponential relationship to linear relationship, which however misuses exponential distance parameter to formulate Gaussian density function. Hathaway [33] gave a general interpretation that the EM algorithm of GMMs is a penalized version of the hard means clustering algorithm. Ichihashi et al. [34] proposed a modified version of FCM with regularization by Kullaback–Leibler (KL) information in the fuzzy objective function. The relation and comparison between KLFCM and GMMs are

* Corresponding author.
E-mail addresses: juzhaojie@hotmail.com, zhaojie.ju@port.ac.uk (Z. Ju).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.08.028

discussed in [34,35]. As a KLFCM algorithm variant, a generic methodology for finite mixture model fitting under a fuzzy clustering principle is proposed and applied to three types of finite mixture models in [36] to improve their performances.

In order to integrate the advantages of GMMs and FCM in the mathematical modelling, conventional GMMs are firstly generalized in such a way that generalized Gaussian model is equipped with non-linearity and a better performance, and then Fuzzy Gaussian Mixture Models (FGMMs) are proposed based on the generalized GMMs for a much faster convergence process which makes it more practical. The dissimilarity function in FCM maintaining the exponential relationship between membership and distance is refined for FGMMs with a degree of fuzziness in terms of the membership grades. Therefore, FGMMs not only possess non-linearity but also have a computationally inexpensive convergence process, which is testified by experiments comparing FGMMs with conventional GMMs and generalized GMMs in terms of fitting degree and convergence speed. The FGMMs are different from the type-2 FGMMs, which uses interval likelihoods to describe the observation uncertainty [37]. The type-2 FGMM is focused on the role of footprint of uncertainty in pattern classification to handle GMMs uncertain mean vector or uncertain covariance matrix, while the FGMM in this paper focuses on the precise parameter estimation of GMMs based on the modified FCM algorithm. The aim of this paper is to improve the conventional GMMs in terms of both performance and efficiency. This paper is organized as follows: Section 2 introduces generalized GMMs; Section 3 proposes two types of FGMMs based on the generalized GMMs; Section 4 provides comparison results of the proposed FGMMs with conventional GMMs and generalized GMMs respectively on various kinds of datasets; finally the paper is concluded with remarks and future work.

2. Generalized GMMs

In this section, we extend the conventional GMMs into a generalized version which enables the GMMs to have capability of modelling curve datasets. Following a brief review of conventional GMMs, an EM algorithm is proposed for the generalized GMMs.

2.1. Conventional GMMs

The probability density function for a Gaussian distribution is given by the formula:

$$p(x|\theta)=\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right) \tag{1}$$

where the set of parameters has $\theta=(\mu,\Sigma)$, μ is the mean, Σ is the covariance matrix of the Gaussian, d is the dimension of vector x, and exp denotes the exponential function.

Let $\mathcal{X}=\{x_1,\ldots,x_n\}$ be a d-dimensional observed dataset of n vectors. If the distribution of $\mathcal{X}$ can be modelled by a mixture of k Gaussians, the density of each vector is

$$p(x_t|\Theta)=\sum_{i=1}^{k}\alpha_i p_i(x_t|\theta_i) \tag{2}$$

where the parameters are $\Theta=(\alpha_1,\ldots,\alpha_k,\theta_1,\ldots,\theta_k)$ and $(\alpha_1,\ldots,\alpha_k)$ are the k mixing coefficients of the k mixed components such that $\sum_{i=1}^{k}\alpha_i=1$; each p_i is a density function parameterized by θ_i. The resulting density for the samples is

$$p(\mathcal{X}|\Theta)=\prod_{t=1}^{n}p(x_t|\Theta)=\mathcal{L}(\Theta|\mathcal{X}) \tag{3}$$

The function $\mathcal{L}(\Theta|\mathcal{X})$ is called the likelihood of the parameters given the data, or the likelihood function. The likelihood is considered as a function of the parameters Θ where the data $\mathcal{X}$ is fixed. In the maximum likelihood problem, the objective is to estimate the parameters set Θ that maximizes $\mathcal{L}$. That is to find Θ^* where

$$\Theta^*=\arg\max_{\Theta}\mathcal{L}(\Theta|\mathcal{X}) \tag{4}$$

Usually, the $\log(\mathcal{L}(\Theta|\mathcal{X}))$ is maximized instead because it is analytically easier. The log-likelihood expression is given by

$$\log(\mathcal{L}(\Theta|\mathcal{X}))=\log\left(\prod_{t=1}^{n}p(x_t|\Theta)\right)=\sum_{t=1}^{n}\log\left(\sum_{i=1}^{k}\alpha_i p_i(x_t|\theta_i)\right) \tag{5}$$

Directly maximizing the log-likelihood is difficult, hence an auxiliary objective function Q is taken into account:

$$Q=\sum_{t=1}^{n}\sum_{i=1}^{k}w_{it}\log[\alpha_i p_i(x_t|\theta_i)] \tag{6}$$

where w_{it} is a posteriori probability for individual class i, $i=1,\ldots,k$, and it satisfies

$$w_{it}=\frac{\alpha_i p_i(x_t|\theta_i)}{\sum_{s=1}^{k}\alpha_s p_s(x_t|\theta_s)} \tag{7}$$

and

$$\sum_{i=1}^{k}w_{it}=1 \tag{8}$$

Maximizing Eq. (6) guarantees that $p(\mathcal{X}|\Theta)$ is maximized if it is performed by an EM algorithm (e.g., [38,39]). The iteration of an EM algorithm estimating the new parameters in terms of the old parameters is given as follows:

- E-step: Compute 'expected' classes of all data points for each class using Eq. (7).
- M-step: Compute maximum likelihood given the data's class membership distributions according to the equations

$$\alpha_i^{new}=\frac{1}{n}\sum_{t=1}^{n}w_{it} \tag{9}$$

$$\mu_i^{new}=\frac{\sum_{t=1}^{n}w_{it}x_t}{\sum_{t=1}^{n}w_{it}} \tag{10}$$

$$\Sigma_i^{new}=\frac{\sum_{t=1}^{n}w_{it}(x_t-\mu_i^{new})(x_t-\mu_i^{new})^T}{\sum_{t=1}^{n}w_{it}} \tag{11}$$

When training GMMs, k-means is employed for initialization before EM starts. The iteration of EM algorithm stops when the change value of log-likelihood is below a preset threshold.

2.2. Generalized Gaussian models

The conventional Gaussian model has intrinsic linearity as its axes are all beelines, so more components are needed when fitting datasets with non-linear manifolds. Active curve axis Gaussian model (AcaG) has bent principal axis, which makes it powerful in modelling curve datasets [22].

In this paper, the generalized Gaussian model is defined as the model including two modalities: one is the conventional Gaussian model with linear axes and the other is bent Gaussian or AcaG model with curve principal axis. Let $\mathcal{X}=\{x_1,\ldots,x_n\}$ be a d-dimensional observed dataset of n vectors. The distribution of $\mathcal{X}$ is based on one Gaussian or bent Gaussian.