Optik xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Original research article

Model selection for Gaussian mixture model based on desirability level criterion

Weishi Peng

School of Equipment Engineering, People Armed Police Engineering University, Xi'an, Shaanxi, 710086, China

ARTICLE INFO

Article history:
Received 4 September 2016
Accepted 31 October 2016

Keywords:
Gaussian mixture model
EM algorithm
Model selection
Desirability level criterion

ABSTRACT

The expectation maximization (EM) algorithm is the most enduring way to estimate the parameters of Gaussian mixture models. However, use the EM algorithm needs to know in advance the true number of mixing components. Therefore, unless this key information is available, it is usually not straightforward to perform this algorithm. On the other hand, its performance highly depends on the initial parameters. To alleviate these problems, a new model selection criterion, i.e., the desirability level criterion, is proposed to choose the number of components. In particular, we proposed a variable step until find either coincides with the actual number or slightly exceeds it, which maximize the value of the desirability level criterion that provides an efficient index to quantify the distance between the Gaussian mixture model fits the observation data. Furthermore, unwanted components can be suppressed by setting the threshold of the desirability level criterion. Numerical examples are provided to illustrate the effectiveness of our desirability level criterion.

1. Introduction

Gaussian mixture model (GMM) is a flexible, powerful probabilistic, and well-weathered models of applied include astronomy, biology, genetics, medicine, psychiatry, economics, engineering et al. (see, e.g., [1,4–6,25,29]). In practical applications, the standard tool for estimating the parameters of GMM is the expectation-maximization (EM) algorithm [2], which can converge in finite iterations. Redner and Walker [3] explained this algorithm, along with helpful remarks about its performance in learning mixtures of univariate Gaussians. Since then, the EM algorithm has received a great amount of attention due to its increasing used in the problem of learning GMM.

However, the EM algorithm for Gaussian mixture fitting has some limitations and drawbacks. It is a local greedy method, and its performance highly depends on the initialization. In addition, the true number of Gaussian components is assumed to known, whereas in several cases this key information is not available. Thus, an actual number of Gaussian must be made along with the parameter estimation, which becomes a crucial issue in Gaussian mixture modelling. Generally, an appropriate number of Gaussian can be chosen via some information and statistical selection criteria. Based on information theory, several model selection methods have been proposed to estimate the number of components of a mixture. Such as Akaike's information criterion (AIC) [7] and its extensions [8], Rissanen's minimum description length (MDL) criterion [9], Schwarz's Bayesian inference criterion (BIC) [10], Mclachlan's Laplace-empirical criterion (LEC) [1], the minimum message length (MML) criterion [11], Bozdogan's informational complexity (ICOMP) criterion [12] and its applications (see, e.g. [13,14]), Banfield's approximate weight of evidence (AWE) criterion [15], and Celeux's normalized entropy criterion (NEC)

E-mail address: peng_weishi@163.com

http://dx.doi.org/10.1016/j.ijleo.2016.10.125
0030-4026/© 2016 Elsevier GmbH. All rights reserved.

[16]. Unfortunately, these algorithms require repeating parameter estimation for different components, and thus it produced a huge computational cost of time and space. What's worst, the correct rate is usually low. Thus, several stochastic simulations were presented to infer the mixture model, such as the Markov chain Monte Carlo methods [17] and Dirichlet processes [18], the Variational Bayesians principle [19], and cross-validation approaches [20] have also been used to calculate the number of mixture components. [21,22] presented a greedy learning method to decide the Gaussian mixture components. Though these stochastic criteria are more efficient than the information criteria, their structure are still complex. On the other hand, EM algorithm is a local greedy method, thus its performance highly depends on the initialization. To overcome this problem, a method called random starting was proposed in Ref. [1]. Ueda et al. [23] presented a split and merge operations and Ma and He [24] introduced a Bayesian Ying-Yang (BYY) harmony learning method to escape from local maxima of the log-likelihood. In addition, clustering [15], unsupervised learning method [27], and deterministic annealing [28] has been used to initialization. Recently, binary tree search method [29], stochastic search method [31], energy-based competitive learning (EBCL) [30], the new initialization strategy [32] and random swap EM algorithm [33] have been proposed to solve the model selection and initialization problems. However, the model selection and initialization problem have not been completely solved yet.

In this paper, we try to propose a new model selection criterion for GMM in terms of the features of the given data. That is, the desirability level criterion is proposed to descript the closeness between the new GMM and the histogram of the data. In particular, unwanted components can be suppressed by setting the threshold of the desirability level criterion. Finally, in order to check the new model selection criterion, we construct a variable step greedy EM algorithm.

This paper is organised as follows. The EM algorithm and its previous work on model selection and initialization are reviewed in Section 2. In Section 3, the desirability level criterion is presented to decide the number of the components in the GMM. Numerical examples are provided in Section 4. Section 5 concludes the paper.

2. Summary of the EM algorithm and its model selection criterion

2.1. EM algorithm

As presented in [1], a GMM is a probability density function (PDF) consisting of a weighted sum of Gaussian densities, which is defined as

$$f(e|\Theta) = \sum_{k=1}^{M} \omega_k N(e|\theta_k) \tag{1}$$

where the weights satisfy the following conditions $\sum_{k=1}^{M} \omega_k = 1, \quad \forall \omega_k \geq 0$.
And the component densities in a d-dimensional is

$$N(e|\theta_k) = N(e|\mu_k, \Sigma_k) = (2\pi)^{-d/2} \det(\Sigma_k)^{1/2} \exp((-1/2) \times (e - \mu_k)^T \Sigma_k^{-1} (e - \mu_k))$$

where the mean $\mu_k \in R^d$, and the covariance matrix Σ_k are collectively denoted by the parameter vector $\theta_k = \{ \mu_k; \Sigma_k\}$. Thus, the GMM is specified by the set of parameters $\Theta = \{\omega_1, \cdots, \omega_M; \quad \theta_1, \theta_2, \cdots, \theta_M\}$.

As we well-known, a variety of learning algorithms were presented for estimating the parameters of mixture model with a sample data set, and the most popular one is the expectation-maximization (EM) algorithm, which converges to a maximum likelihood estimate of the mixture parameters. Assume $e = \{e_i\}_{i=1}^{n}$ is the training set. Then, the parameters Θ are estimated via the following iterative update equations for each component $k, \quad k = 1, 2, \cdots, M$:

$$\omega_k^{t+1} = \frac{1}{n} \sum_{i=1}^{n} P(k|e_i) \tag{2}$$

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} P(k|e_i) \times e_i}{\sum_{i=1}^{n} P(k|e_i)} \tag{3}$$

$$\Sigma_k^{t+1} = \frac{\sum_{i=1}^{n} P(k|e_i) \times (e_i - \mu_k^{t+1}) \times (e_i - \mu_k^{t+1})^T}{\sum_{i=1}^{n} P(k|e_i)} \tag{4}$$