Pattern Recognition 45 (2012) 2804–2816

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Maximum likelihood estimation of Gaussian mixture models using stochastic search

Çağlar Arı [a], Selim Aksoy [b,*], Orhan Arıkan [a]

[a] Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey
[b] Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

ARTICLE INFO

Article history:
Received 28 March 2011
Received in revised form
16 December 2011
Accepted 30 December 2011
Available online 11 January 2012

Keywords:
Gaussian mixture models
Maximum likelihood estimation
Expectation–maximization
Covariance parametrization
Identifiability
Stochastic search
Particle swarm optimization

ABSTRACT

Gaussian mixture models (GMM), commonly used in pattern recognition and machine learning, provide a flexible probabilistic model for the data. The conventional expectation–maximization (EM) algorithm for the maximum likelihood estimation of the parameters of GMMs is very sensitive to initialization and easily gets trapped in local maxima. Stochastic search algorithms have been popular alternatives for global optimization but their uses for GMM estimation have been limited to constrained models using identity or diagonal covariance matrices. Our major contributions in this paper are twofold. First, we present a novel parametrization for arbitrary covariance matrices that allow independent updating of individual parameters while retaining validity of the resultant matrices. Second, we propose an effective parameter matching technique to mitigate the issues related with the existence of multiple candidate solutions that are equivalent under permutations of the GMM components. Experiments on synthetic and real data sets show that the proposed framework has a robust performance and achieves significantly higher likelihood values than the EM algorithm.

1. Introduction

Gaussian mixture models (GMMs) have been one of the most widely used probability density models in pattern recognition and machine learning. In addition to the advantages of parametric models that can represent a sample using a relatively small set of parameters, they also offer the ability of approximating any continuous multi-modal distribution arbitrarily well like nonparametric models by an appropriate choice of its components [1,2]. This flexibility of a convenient semiparametric nature has made GMMs a popular choice for both density models in supervised classification and cluster models in unsupervised learning problems.

The conventional method for learning the parameters of a GMM is maximum likelihood estimation using the expectation–maximization (EM) algorithm. Starting from an initial set of values, the EM algorithm iteratively updates the parameters by maximizing the expected log-likelihood of the data. However, this procedure has several issues in practice [1,2]. One of the most important of these issues is that the EM algorithm easily gets trapped in a local maximum as the objective being a non-concave optimization problem. Moreover, there is also the associated problem of initialization as it influences which local maximum of the likelihood function is attained.

The common approach is to run the EM algorithm many times from different initial configurations and to use the result corresponding to the highest log-likelihood value. However, even with some heuristics that have been proposed to guide the initialization, this approach is usually far from providing an acceptable solution especially with increasing dimensions of the data space. Furthermore, using the results of other algorithms such as *k*-means for initialization is also often not satisfactory because there is no mechanism that can measure how different these multiple initializations are from each other. In addition, this is a very indirect approach as multiple EM procedures that are initialized with seemingly different values might still converge to similar local maxima. Consequently, this approach may not explore the solution space effectively using multiple independent runs.

Researchers dealing with similar problems have increasingly started to use population-based stochastic search algorithms where different potential solutions are allowed to interact with each other. These approaches enable multiple candidate solutions to simultaneously converge to possibly different optima by making use of the interactions. Genetic algorithm (GA) [3–7], differential evolution (DE) [8], and particle swarm optimization (PSO) [9–12] have been the most common population-based stochastic search algorithms used for the estimation of some form of GMMs. Even though these approaches have been shown to perform better than non-stochastic alternatives such as *k*-means and fuzzy *c*-means, the interaction

* Corresponding author. Tel.: +90 312 2903405; fax: +90 312 2664047.
E-mail addresses: cari@ee.bilkent.edu.tr (Ç. Arı), saksoy@cs.bilkent.edu.tr (S. Aksoy), oarikan@ee.bilkent.edu.tr (O. Arıkan).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.12.023

mechanism that forms the basis of the power of the stochastic search algorithms has also limited the use of these methods due to some inherent assumptions in the candidate solution parametrization. In particular, the interactions in the GA, DE, and PSO algorithms are typically implemented using randomized selection, swapping, addition, and perturbation of the individual parameters of the candidate solutions. For example, the crossover operation in GA and DE randomly selects some parts of two candidate solutions to create a new candidate solution during the reproduction of the population. Similarly, the mutation operation in GA and DE and the update operation in PSO perturb an existing candidate solution using a vector that is created using some combination of random numbers and other candidate solutions. However, randomized modification of individual elements of a covariance matrix independently does not guarantee the result to be a valid (i.e., symmetric and positive definite) covariance matrix. Likewise, partial exchanges of parameters between two candidate solutions lead to similar problems. Hence, these problems confined the related work to either use no covariance structure (i.e., implicitly use identity matrices centered around the respective means) [7–10,12] or constrain the covariances to be diagonal [3,11]. Consequently, most of these approaches were limited to the use of only the mean vectors in the candidate solutions and to the minimization of the sum of squared errors as in the *k*-means setting instead of the maximization of a full likelihood function. Full exploitation of the power of GMMs involving arbitrary covariance matrices estimated using stochastic search algorithms benefits from new parametrizations where the individual parameters are *independently modifiable* so that the resulting matrices remain valid covariance matrices after the stochastic updates and have finite limits so that they can be searched within a *bounded* solution space. In this paper, we present a new parametrization scheme that satisfies these criteria and allows the estimation of generic GMMs with arbitrary covariance matrices.

Another important problem that has been largely ignored in the application of stochastic search algorithms to GMM estimation problems in the pattern recognition literature is identifiability. In general, a parametric family of probability density functions is identifiable if distinct values of the parameters determine distinct members of the family [1,2]. For mixture models, the identifiability problem exists when there is no prior information that allows discrimination between its components. When the component densities belong to the same parametric family (e.g., Gaussian), the mixture density with K components is invariant under the $K!$ permutations of the component labels (indices). Consequently, the likelihood function becomes invariant under the same permutation, and this invariance leads to $K!$ equivalent modes, corresponding to equivalence classes on the set of mixture parameters. This lack of *uniqueness* is not a cause for concern for the iterative computation of the maximum likelihood estimates using the EM algorithm, but can become a serious problem when the estimates are iteratively computed using simulations when there is the possibility that the labels (order) of the components may be switched during different iterations [1,2]. Considering the fact that most of the search algorithms depend on the designed interaction operations, performances of the operations that assume continuity or try to achieve diversity cannot work as intended, and the discontinuities in the search space will make it harder for the search algorithms to find directions of improvement. In an extreme case, the algorithms will fluctuate among different solutions in the same equivalence class, hence, among several equivalent modes of the likelihood function, and will have significant convergence issues. In this paper, we propose an optimization framework where the optimal correspondences among the components in two candidate solutions are found so that desirable interactions become possible between these solutions.

It is clear that a formulation that involves *unique*, *independently modifiable*, and *bounded* parameters is highly desired for effective utilization of stochastic search algorithms for the maximum likelihood estimation of unrestricted Gaussian mixture models. Our major contributions in this paper are twofold: we present a novel parametrization for arbitrary covariance matrices where the individual parameters can be independently modified in a stochastic manner during the search process, and describe an optimization formulation for resolving the identifiability problem for the mixtures. Our first contribution, the parametrization, uses eigenvalue decomposition, and models a covariance matrix in terms of its eigenvalues and Givens rotation angles extracted using QR factorization of the eigenvector matrices via a series of Givens rotations. We show that the resulting parameters are independently modifiable and are bounded so they can be naturally used in different kinds of stochastic global search algorithms. We also describe an algorithm for ordering the eigenvectors so that the parameters of individual Gaussian components are uniquely identifiable.

As our second major contribution, we propose an algorithm for ordering of the Gaussian components within a candidate solution for obtaining a unique correspondence between two candidate solutions during their interactions for parameter updates throughout the stochastic search. The correspondence identification problem is formulated as a minimum cost network flow optimization problem where the objective is to find the correspondence relation that minimizes the sum of Kullback–Leibler divergences between pairs of Gaussian components, one from each of the two candidate solutions. We illustrate the proposed parametrization and identifiability solutions using PSO for density estimation. An early version of this paper [13] presented initial experiments on clustering.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 establishes the notation and defines the estimation problem. Section 4 summarizes the EM approach for GMM estimation. Section 5 presents the details of the proposed covariance parametrization and the solution for the identifiability problem. Section 6 describes the PSO framework and its adaptation as a stochastic search algorithm for GMM estimation. Section 7 presents the experiments and discussion using both synthetic and real data sets. Finally, Section 8 provides the conclusions of the paper.

2. Related work

As discussed in the previous section, existing work on the use of stochastic search algorithms for GMM estimation typically uses only the means [7–10,12] or means and standard deviations alone [3,11] in the candidate solutions. Exceptions where both mean vectors and full covariance matrices were used include [4,5] where EM was used for the actual local optimization by fitting Gaussians to data in each iteration and GA was used only to guide the global search by selecting individual Gaussian components from existing candidate solutions in the reproduction steps. However, treating each Gaussian component as a whole in the search process and fitting it locally using the EM iterations may not explore the whole solution space effectively especially in higher dimensions. Another example is [6] where two GA alternatives for the estimation of multidimensional GMMs were proposed. The first alternative encoded the covariance matrices for d-dimensional data using $d+d^2$ elements where d values corresponded to the standard deviations and d^2 values represented a correlation matrix. The second alternative used d runs of a GA for estimating 1D GMMs followed by d runs of EM starting from the results of the GAs. Experiments using 3D synthetic data