Egyptian Informatics Journal (2011) **12**, 19–27

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

An algorithm for unsupervised learning and optimization of finite mixture models

Ahmed R. Abas

Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Egypt

Received 4 May 2010; accepted 28 September 2010
Available online 22 March 2011

KEYWORDS

Finite Mixture Models;
Expectation–Maximization;
Unsupervised learning;
Clustering;
Optimization

Abstract In this paper, an algorithm is proposed to integrate the unsupervised learning with the optimization of the Finite Mixture Models (**FMM**). While learning parameters of the **FMM** the proposed algorithm minimizes the mutual information among components of the **FMM** provided that the reduction in the likelihood of the **FMM** to fit the input data is minimized. The performance of the proposed algorithm is compared with the performances of other algorithms in the literature. Results show the superiority of the proposed algorithm over the other algorithms especially with data sets that are sparsely distributed or generated from overlapped clusters.

1. Introduction

Unsupervised learning or cluster analysis is an important task in pattern recognition. It is interested in grouping similar feature vectors in an input data set into a number of groups or clusters. Feature vectors belonging to the same cluster are similar to each other more than to other feature vectors belonging to the other clusters. Several clustering algorithms are proposed in the literature such as the **K-means** algorithm, and the **FMM** [1,2]. The **FMM** is preferred for cluster analysis because it produces a certainty estimate of the membership of each feature vector to each one of the clusters in the input data set. Each component in the **FMM** is usually a Gaussian distribution. Unsupervised learning of the **FMM** parameters is usually achieved via the Expectation–Maximization (**EM**) algorithm [3]. The **EM** algorithm determines the **FMM** parameters that maximize the likelihood of this **FMM** to fit the input data set. However, the **EM** algorithm has some limitations. First, it produces sub-optimal results as it converges to the nearest local maximum of the likelihood function to the starting point. Second, it produces biased estimates for the mixture parameters when clusters are poorly separated i.e., overlapped, or when mixing weights of the mixture components have extreme values i.e., data are sparsely distributed [4]. Optimization of a **FMM** is defined as the minimization of the number of components in the **FMM** required for fitting an input data

E-mail address: arabas@zu.edu.eg

Peer review under responsibility of Faculty of Computers and Information, Cairo University.
doi:10.1016/j.eij.2011.02.005

Production and hosting by Elsevier

set. Optimization is one of the most difficult problems in cluster analysis [5].

Several criteria are proposed in the literature for the estimation of the number of **FMM** components and hence the number of clusters assuming that each cluster is represented by a component in the **FMM**. A group of these criteria is the penalized-likelihood criteria, which include as examples the Bayesian Information Criterion (**BIC**) [6], the Bezdek's Partition Coefficient (**PC**) [7], and the Minimum Message Length (**MML**) criterion [8]. Other examples are the Information Theoretic Measure of Complexity (**ICOMP**) [9,10], the Minimum Description Length (**MDL**) criterion [11], the Akaike's Information Criterion (**AIC**) [12], the Approximate Weight of Evidence (**AWE**) criterion [13], and the Evidence-Based Bayesian (**EBB**) criterion [14]. Also, a new **MML**-like criterion is proposed [15] and used with the Component-Wise EM (**CEM**) algorithm [16] to estimate the number of **FMM** components. The resulting algorithm overcomes problems of the common **EM** algorithm such as obtaining sub-optimal results; and approaching the boundary of the parameter space when at least one of the components becomes too small. However, due to the dependency on the **EM** algorithm the model selected using these criteria is not necessarily the best model for clustering small data sets. In other words, the selected model does not necessarily represent well-separated clusters that are clearly associated with the model components [17]. It has been shown that the **BIC/MDL** criterion performs comparably with both of the **EBB** and the **MML** criteria, and it outperforms many other criteria in the literature [14]. The **BIC/MDL** criterion has been shown to produce a good approximation to Bayes factor [18]. However, although the **BIC/MDL** criterion is preferred when data clusters are separated and the data size is large [19], it tends to overestimate the number of components when cluster shapes are not Gaussian [4]. On the other hand, it tends to underestimate the number of components when clusters are overlapped or when the number of feature vectors in the given data set is small [20]. Penalized-likelihood criteria compromise the goodness of fitting of the **FMM** to the input data set with the complexity of that **FMM**. Since the mixture complexity is a quadratic function of the number of features (dimensions) in the input data set these criteria are sensitive to the increase of the number of features in the input data set. In the rest of this paper, the algorithms that use the **BIC** and the **MML** criteria for determining the number of **FMM** components are referred to as the **BIC** algorithm and the **MML** algorithm, respectively.

Another group of criteria for the estimation of the number of **FMM** components is based on the mutual information. This group includes Data Entropy that is used to evaluate different mixture models with different number of components [21]. However, this criterion may overestimate the number of components in the presence of outliers, as it is biased toward producing separated components. Another criterion in this group based on the Bayesian-Kullback Ying-Yang learning theory [22] is proposed [23]. This criterion is used in determining the number of **FMM** components [5]. However, due to the dependency on the **EM** algorithm for learning mixture model parameters this criterion has the same drawbacks of the penalized-likelihood criteria. Therefore, this criterion produces inaccurate results with small data sets [5]. Also, an algorithm that is based on the mutual information theory is proposed [20]. However, on the opposite of the algorithms that use the penalized-likelihood criteria, this algorithm removes the largest component that is overlapped with many other small components in the **FMM**. This results in bad quality of the cluster structure obtained by the resulting **FMM** because large components in the **FMM** are supported by the data more than small components. In addition, deleting large components in the **FMM** causes the likelihood function to be largely decreased. This algorithm also underestimates the number of mixture components when some clusters are poorly separated in the data space. Finally, the authors used only centers of the mixture components instead of all the data points in their definition of the mutual information between two components in the **FMM**. This may be only valid with data sets that are dense and concentrated around their cluster centers as the examples shown by the authors. In the rest of this paper, this algorithm is referred to as the Mutual Information (**MI**) algorithm. Another algorithm that is based on mutual information theory is proposed [24]. However, this algorithm has initialization problem due to starting with small number of components in the mixture model. In addition, this algorithm has satisfactory results in determining the number of mixture components that is equal to the number of clusters of the input data set only when the size of this data set is large as reported by the authors. With small data sets, especially those data sets that are sparsely distributed and generated from overlapped clusters, this algorithm underestimates the number of mixture components due to the use of the histogram method for density estimation. Recently, a Bayesian Ying–Yang (**BYY**) scale-incremental **EM** algorithm for Gaussian mixture learning for both the parameter estimation and model selection is proposed [25]. However, this algorithm has initialization problem due to starting with small number of components in the mixture model and using the **BYY** harmony function as a stopping criterion that depends on the estimated values of mixture parameters via the **EM** algorithm. In addition, with small data sets, especially those data sets that are sparsely distributed and generated from overlapped clusters, this algorithm underestimates the number of mixture components because the **BYY** harmony function is biased toward producing well separated clusters of nearly equal size.

Different criteria for the estimation of the number of **FMM** components include Adaptive Mixtures algorithm that is a recursive form of the **EM** algorithm [26]. Although this algorithm does not require a range of the number of components, it may overestimate the number of components when the given data set contains sparsely distributed data [20]. Also, it may underestimate the number of components when some clusters in the data space are poorly separated. This results from the iterative form of the **EM** algorithm, which may generate an unnecessary component for few outliers in the data set and also may allow many components to be overlapped. In addition, the resulting model depends on the order of presenting the input data patterns to the algorithm due to the recursive nature of the algorithm. Finally, this algorithm does not have a measure that compromises the increase in the **FMM** complexity with the goodness of fitting of that model to the given data. A cross-validated likelihood criterion is proposed to estimate the number of components in the **FMM** using large data sets [27]. However, this criterion requires not only a large data set in order to be divided into training and test data but also a sufficient range of the number of components. In addition, the selected model is not necessarily the optimum