Efficient greedy estimation of mixture models through a binary tree search

Nicola Greggio [a,*], Alexandre Bernardino [a], Paolo Dario [b], José Santos-Victor [a]

[a] *Instituto de Sistemas e Robótica, Instituto Superior Técnico, 1049-001 Lisboa, Portugal*
[b] *BioRobotics Institute, Scuola Superiore Sant'Anna, 56127 Pisa, Italy*

HIGHLIGHTS

- Unsupervised estimation of the mixture's parameters and complexity, simultaneously.
- Important computational savings with respect to alternative approaches.
- Unique initialization procedure, allowing a deterministic evolution of the algorithm.
- Mixture parameters' estimation with a modified Expectation Maximization (EM) algorithm.
- MML information criteria for comparing models with different complexity.

ARTICLE INFO

ABSTRACT

Unsupervised data clustering can be addressed by the estimation of mixture models, where the mixture components are associated to clusters in data space. In this paper we present a novel unsupervised classification algorithm based on the simultaneous estimation of the mixture's parameters and the number of components (complexity). Its distinguishing aspect is the way the data space is searched. Our algorithm starts from a single component covering all the input space and iteratively splits components according to breadth first search on a binary tree structure that provides an efficient exploration of the possible solutions. The proposed scheme demonstrates important computational savings with respect to other state-of-the-art algorithms, making it particularly suited to scenarios where the performance time is an issue, such as in computer and robot vision applications. The initialization procedure is unique, allowing a deterministic evolution of the algorithm, while the parameter estimation is performed with a modification of the Expectation Maximization algorithm. To compare models with different complexity we use the Minimum Message Length information criteria that implement the trade-off between the number of components and data fit log-likelihood. We validate our new approach with experiments on synthetic data, and we test and compare to related approaches its computational efficiency in data-intensive image segmentation applications.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Data clustering is a topic of major interest in many disciplines, having an extremely wide application range. Several techniques for clustering have been developed during the last decades, including Kohonen maps [1,2], Growing Neural gas [3,4], k-means [5], to Independent component analysis [6,7], and mixture models [8].

To cope with the dynamics of changing environments, robotics applications must process in real-time huge amounts of sensory data. For example, image processing algorithms play a key role in many practical applications and demand significant computational resources due to huge amounts of data.

A paradigmatic application of clustering in robot vision is image segmentation. Several applications in humanoid robots [9,10], rescue robots [11], or soccer robots [12] rely on some sort of image segmentation [13]. Additionally, many other fields of image analysis depend on image segmentation algorithms: video surveillance, medical imaging and database retrieval are some examples [14,15].

In this paper we propose an unsupervised data clustering algorithm for data-intensive algorithms, moving towards practical applications in robotics and computer vision. We propose computational efficiency improvements to the state-of-the-art

* Corresponding author.
E-mail addresses: nicola.greggio@sssup.it, ngreggio@isr.ist.utl.pt (N. Greggio).

unsupervised clustering algorithms and illustrate the achieved performances and computational gains both in synthetic data and on image segmentation applications. Our goal is not to improve the state of the art in image segmentation, but to demonstrate efficiency gains with respect to alternative approaches. The state of the art in image segmentation technology is much beyond simple data clustering, but our methodology can certainly be useful for segmentation techniques requiring clustering algorithms.

Our approach to data clustering is based on the Expectation Maximization algorithm applied to Gaussian mixtures, which allows us to approximate arbitrary probability distribution functions. Fitting a mixture model to the distribution of the given data is equivalent, in some applications, to the identification of the clusters with the mixture components [8]. The selection of the right model complexity, i.e. the number of mixture components, is a critical issue. In fact, this is an ill-posed problem with multiple possible solutions depending on a trade-off between data fit and complexity. The higher the number of components in the mixture is, the better the data fit will be. Unfortunately, increasing the number of components will lead both to data over fitting and to increase in the computational burden. Therefore, finding the best compromise between precision, generalization and speed is an essential concern. A common approach to address this compromise is to try different hypotheses for the number of components in the mixture, and then selecting the best one according to some appropriate model selection criteria. In such approaches, the best model is typically selected by executing independent runs of the EM algorithm for many different initializations of the number of components, and then evaluating each of the solutions with criteria that penalize complex models. Some examples are the Akaike Information Criterion (AIC) [16], the Schwarz's Bayesian Information Criterion [17], the Rissanen Minimum Description Length (MDL) [18], and Wallace and Freeman Minimum Message Length (MML) [19]. However, all of these criteria, in order to be effective, have to be evaluated for every possible number of models under comparison. Thus, it is a combinatorial problem where there is no guarantee of finding an optimal global solution, except with exhaustive search methods.

1.1. Our contribution

Our work follows an incremental model selection approach. Instead of executing independent EM runs spanning a wide range of parameters, and evaluating each one with a model selection criteria, we change the model complexity on-the-fly, while a modified EM algorithm optimizes the mixture parameters. The search for the best number of components in the mixture starts with a single component and progressively adds new components according to a binary tree structure, i.e. forking one of the existing components. Using a breadth-first search scheme on the binary tree structure allows an efficient exploration of the search space and provides "backtracking" capabilities, i.e. if the addition of a new element does not prove to be useful, we can go back to the best previous solution and continue exploration on a different part of the search space. Such a strategy allows a good search efficiency allied to a good coverage of the possible solution's space, achieving better fits than alternative algorithms, in average. The current paper is an extension of the contents presented in [20]. We compare the proposed method both with our previous work [21] and with the method in Figueiredo and Jain [22].

2. Related work

Due to the complexity of the model selection problem, many algorithms have been proposed to escape the classical exhaustive search methods by adopting an incremental approach. Most of these algorithms derive from the original EM formulation, but they are capable of modifying their complexity during the learning process. Here, it is possible to distinguish three different categories: those starting with a low number of components and only increase their number as the algorithm progresses; those starting with a high number and annihilating components along time; and those both incrementing and reducing the number of components. Typically new components are added by random initialization (birth operator) or by dividing existing components (split operation), whereas components are removed either by annihilation (death operator) or joining other existing components (merge operator). Split-and-merge algorithms have been widely used in computer vision, pattern recognition and signal processing [23,24], and they are often more efficient than exhaustive, random or genetic algorithm approaches. Richardson and Green used split-and-merge operations together with birth-and-death operations to develop a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm for fully Bayesian analysis of univariate Gaussian mixtures [25]. The novel RJMCMC methodology elaborated by Green is attractive because it can preferably deal with parameter estimation and model selection jointly in a single paradigm. BrieGy, a random-sweep Metropolis Hastings method, constructs the dimension matching transform with the reversible jump methodology [26]. Then, Ueda et al. proposed a split-and-merge EM algorithm (SMEM) to alleviate the problem of local convergence of the EM method [27]. They analysed the implications of merge and split operations in terms of their "well-posedness". While merging two components is trivial (the point set in the merged component is the union of the original ones), there are many possible ways to split a component, which results in an ill-posed problem. Merge-only approaches, thus, alleviate this problem. In this class of methods, the approach proposed by Figueiredo and Jain in 2002 is especially interesting [22]. They start with a large number of components and impose a Dirichlet prior on the mixing weights that drives to zero the weights of the components that get low support from data during the steps of the algorithms. Huang et al. proposed a refinement of the method using a new penalized likelihood method that is continuous when components are annihilated and shows good analytic properties [28]. Despite the remarkable accuracy and robustness of merge-only methods, they suffer from a high computational burden given that many EM iterations are performed with a large number of components.

With the aim of reducing as much as possible the required computational demands, split-only approaches start with a single component and add new ones along time according to different criteria. In 2002 Vlassis and Likas introduced a greedy algorithm [29] that starts with a single component covering all the data and sequentially splits it in two new ones. The parameters of these two components are then adjusted by local EM iterations. The method takes $O(n^2)$ operations, where n is the number of input data samples. Subsequently, Verbeek et al. developed another greedy method [30] by starting 'partial' EM searches, each of them with different initializations. The total complexity for the algorithm to learn a sequence of mixtures composed of k components is $O(k^2 n)$, where n is as before.

In our previous work [21] we proposed a split-only method whose decisions were made in accordance to the state of the Gaussian components and a set of adaptive thresholds. In this paper we propose an approach that introduces significant improvements in reducing the number of tuning parameters and performing a better exploration of the search space, leading to a more efficient, robust, and easy to use methodology.

3. Mixture learning algorithm

Due to the nature of the mixture learning problem with model selection, finding globally optimal solutions is very hard except