Semi-Supervised Clustering: Application to Image Segmentation

Mário A.T. Figueiredo

Instituto de Telecomunicações and Instituto Superior Técnico, Technical University of Lisbon, 1049-001 Lisboa, Portugal; mario.figueiredo@lx.it.pt

Abstract. This paper describes a new approach to semi-supervised model-based clustering. The problem is formulated as penalized logistic regression, where the labels are only indirectly observed (via the component densities). This formulation allows deriving a generalized EM algorithm with closed-form update equations, which is in contrast with other related approaches which require expensive Gibbs sampling or suboptimal algorithms. We show how this approach can be naturally used for image segmentation under spatial priors, avoiding the usual hard combinatorial optimization required by classical Markov random fields; this opens the door to the use of sophisticated spatial priors (such as those based on wavelet representations) in a simple and computationally very efficient way.

1 Introduction

In recent years there has been a considerable amount of interest in semi-supervised learning problems (see Zhu (2006)). Most formulations of semi-supervised learning approach the problem from one of the two ends of the unsupervised-supervised spectrum: either supervised learning in the presence of unlabelled data (see, e.g., Belkin and Niyogi (2003), Krishnapuram et al. (2004), Seeger (2001), Zhu et al. (2003)) or unsupervised learning with additional information (see, e.g., Basu et al. (2004), Law et al. (2005), Lu and Leen (2005), Shental et al. (2003), Wagstaff et al. (2001)). The second perspective, known as semi-supervised clustering (SSC), is usually adopted when labels are completely absent from the training data, but there are (say, pair-wise) relations that one wishes to enforce or simply encourage.

Most methods for SSC work by incorporating the desired relations (or constraints) into classical algorithms such as the expectation-maximization (EM) algorithm for mixture-based clustering or the K-means algorithm. These relations may be imposed in a hard way, as constraints (Shental et al. (2003),

Wagstaff et al. (2001)), or used to build priors under which probabilistic clustering is performed (Basu et al. (2004), Lu and Leen (2005)). This last approach has been shown to yield good results and is the most natural for applications where one knows that the relations should be encouraged, but not enforced (e.g., in image segmentation, neighboring pixels should be encouraged, but obviously not enforced, to belong to the same class). However, the resulting EM-type algorithms have a considerable drawback: because of the presence of the prior on the grouping relations, the E-step no longer has a simple closed form, requiring the use of expensive stochastic (e.g., Gibbs) sampling schemes (Lu and Leen (2005)) or suboptimal methods such as the *iterated conditional modes* (ICM) algorithm (Basu et al. (2004)).

In this paper, we describe a new approach to semi-supervised mixture-based clustering for which we derive a simple, fully deterministic *generalized EM* (GEM) algorithm. The keystone of our approach is the formulation of semi-supervised mixture-based clustering as a penalized logistic regression problem, where the labels are only indirectly observed. The linearity of the resulting complete log-likelihood, with respect to the missing group labels, will allow deriving a simple GEM algorithm.

We show how the proposed formulation is used for image segmentation under spatial priors which, until now, were only used for real-valued fields (e.g., image restoration/denoising): Gaussian fields and wavelet-based priors. Under these priors, our GEM algorithm can be implemented very efficiently by resorting to fast Fourier or fast wavelet transforms. Our approach completely avoids the combinatorial nature of standard segmentation methods, which are based on Markov random fields of discrete labels (see Li (2001)).

Although we focus on image segmentation, SSC has been recently used in other areas, such as clustering of image databases (see Grira et al. (2005)), clustering of documents (see Zhong (2006) for a survey), and bioinformatics (see, e.g., Nikkilä et al. (2001), Cebron and Berthold (2006)). Our approach will thus also be potentially useful in those application areas.

2 Formulation

We build on the standard formulation of finite mixtures: let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be an observed data set, with each $\mathbf{x}_i \in I\!R^d$ assumed to have been generated (independently) according to one of a set of K probability densities $\{p(\cdot|\boldsymbol{\phi}^{(1)}), ..., p(\cdot|\boldsymbol{\phi}^{(K)})\}$. Associated with $\mathcal{X}$, there's a hidden/missing label set $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$, where $\mathbf{y}_i = [y_i^{(1)}, ..., y_i^{(K)}]^T \in \{0,1\}^K$, with $y_i^{(k)} = 1$ if and only if $\mathbf{x}_i$ was generated by source k ("1-of-K" binary encoding). Thus,

$$p\left(\mathcal{X} \left| \mathcal{Y}, \boldsymbol{\phi}^{(1)}, ..., \boldsymbol{\phi}^{(K)} \right.\right) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p(\mathbf{x}_i|\boldsymbol{\phi}^{(k)})\right]^{y_i^{(k)}} . \tag{1}$$