Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 80 – 86

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Variational Inference for Acoustic Unit Discovery

Lucas Ondel*, Lukaš Burget, Jan Černocký

Brno University of Technology, Czech Republic

Abstract

Recently, several nonparametric Bayesian models have been proposed to automatically discover acoustic units in unlabeled data. Most of them are trained using various versions of the Gibbs Sampling (GS) method. In this work, we consider Variational Bayes (VB) as alternative inference process. Even though VB yields an approximate solution of the posterior distribution it can be easily parallelized which makes it more suitable for large database. Results show that, notwithstanding VB inference is an order of magnitude faster, it outperforms GS in terms of accuracy.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords:

Bayesian non-parametric, Variational Bayes, acoustic unit discovery

1. Introduction

Whereas Automatic Speech Recognition (ASR) systems are more and more frequently used in daily life applications, the need of labeled data has never been so high. With the ever-growing use of Internet a huge amount of unlabeled audio data coming from many different countries is now available. However, because the labeling process by human expert is expensive this data has still been unexploited. In [1], a nonparametric Bayesian model to automatically segment and label audio data has been proposed. The model has been later extended in [2] to jointly learn the phonetic units and the word pronunciations. An attempt to tackle the problem by mean of neural networks as also been investigated in [3]. In [1] and [2], both models are trained with the Gibbs Sampling (GS) algorithm [4] which can be summarized as follows: sample a new value for each parameter of the model from the probability of the parameter given the data and the others parameters and repeat until convergence. This method has many advantages. First, it allows the optimization of complex Bayesian model without the need of analytical solutions. Second, it can be shown that the dynamic converges toward the optimal distribution independently of the initial value of the parameters. However, despite these powerful features the algorithm carries some severe drawbacks: the parameters of the model cannot be

* Corresponding author.
E-mail address: iondel@fit.vutbr.cz

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.033

sampled asynchronously and the rate of convergence may be slow. Hence, even though GS is a popular way to train Bayesian models, its power is limited by its inability to handle large amount of data.

The Variational Bayesian (VB) inference ([5]) is an alternative technique to train Bayesian models which copes with the weaknesses of the GS algorithm. In essence, the application of VB is very similar to the well known Expectation-Maximization (EM) algorithm for latent models, allowing the training to be parallelized and the convergence monitored by a lower bound on the likelihood of the model. However, these benefits have a cost: the procedure may converge toward a local optimum. In this work we compare both training algorithm on a model similar to the one presented in[1]. Results show that for a considerable gain of speed the VB training achieves higher accuracy than the GS training.

2. Model

2.1. Model definition

Our model aims at segmenting and clustering unlabeled speech data into phone-like categories. It is similar to a phone-loop model in which each phone-like unit is modeled by an HMM [1]. This phone-loop model is fully Bayesian in the sense that:

- it incorporates a prior distribution over the parameters of the HMMs
- it has a prior distribution over the units modeled by a Dirichlet process[6].

Informally, the Dirichlet process prior can be seen as a standard Dirichlet distribution prior for a Bayesian mixture with an infinite number of components. However, we assume that our N data samples have been generated with only M components ($M \leq N$) from the infinite mixture. Hence, the model is no longer restricted to have a fixed number of components but instead can learn its complexity (i.e. number of components used M) according to the training data. The generation of a data set with M speech units can be summarized as follows:

1. sample the vector $\mathbf{v} = v_1, ..., v_M$ with
$$v_i \sim \text{Beta}(1, \gamma)$$
where γ is the concentration parameters of the Dirichlet process
2. sample M HMM parameters θ_1,...,θ_M from the base distribution of the Dirichlet process.
3. sample each segment as follows:
   (a) choose a HMM parameters with probability $\pi_i(\mathbf{v})$ defined as:
$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$$
   (b) sample a path $\mathbf{s} = s_1, ..., s_n$ from the HMM transition probability distribution
   (c) for each s_i in $\mathbf{s}$:
      i. choose a Gaussian components from the mixture model
      ii. sample a data point from the Gaussian density function

A similar model have been applied in[1], however, two major differences should be noted: first, we have chosen to consider the stick-breaking construction[7] of the Dirichlet process (step 1 and 2 of the generation) rather than the Chinese Restaurant Process (CRP). See[8] and[1] for training Bayesian models with the CRP. This allow us to use variational methods to infer the distribution over the parameters rather than sampling methods. Secondly, our model

[1] For the sake of readability we write HMM for the complete HMM/GMM model.