Pattern Recognition Letters 35 (2014) 236–245

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Handling signal variability with contextual markovian models

Mathieu Radenen *, Thierry Artières

Université Pierre et Marie Curie, LIP6, 4 place Jussieu, 75005 Paris, France

ARTICLE INFO

Article history:
Available online 6 September 2013

Keywords:
Hidden Markov models
Conditional Random Fields
Sequence classification
Robustness to variability

ABSTRACT

There are two popular families of statistical models for dealing with sequences and in particular with handwriting signals, either on-line or off-line, the well known generative hidden Markov models and the more recently proposed discriminative Hidden Conditional Random Fields.

One key issue in such modeling frameworks is to efficiently handle variability. The traditional approach consists in first removing as much as possible signal variability in the preprocessing stage, and to use more complex models, for instance in the case of hidden Markov models one increases the number of states and the Gaussian mixture size.

We focus here on another kind of approaches where the probability distribution implemented by the models depends on a number of additional *contextual variables*, that are assumed fixed or that vary slowly along a sequence. The context may stand for emotion features in speech recognition, physical features in gesture recognition, gender, age, etc.

We propose a framework for deriving markovian models that make use of such contextual information. This yields new models that we call Contextual hidden Markov models and contextual Hidden Conditional Random Fields. We detail learning algorithms for both models and investigate their performances on the IAM off-line handwriting dataset.

1. Introduction

Today hidden Markov models (HMMs) and Hidden Conditional Random Fields (HCRFs) are the two most popular approaches for building systems for sequence modeling, sequence recognition and sequence labeling in a variety of application fields.

HMMs are a famous class of probabilistic generative models that are well known for their efficient algorithms, their simplicity and their robustness for classifying and labeling sequences. Despite their popularity, they rely on strong assumptions and exhibit at the end a limited expressive power. Also HMMs are generative models that were originally trained through Maximum Likelihood Estimation (MLE). This non discriminative training criterion fits well modeling applications where one wants to learn accurate models of production, e.g. for synthesis (Keiichi et al., 2000; Hofer et al., 2007), but it is not well adapted to recognition tasks.

To build more accurate sequence recognition and labeling HMM systems, a first bunch of methods have been proposed to train HMMs in a discriminative way. The most famous approaches are Maximum Mutual Information (Bahl et al., 1986), Minimum Classification Error (Juang and Katagiri, 1992) and Minimum Phone Error named after its initial application to speech recognition (Povey and Woodland, 2002). More recently, few works have applied the large margin principle to HMM learning, most of these works have concerned speech recognition (Sha and Saul, 2007; Cheng et al., 2009; McDermott et al., 2009) and handwriting recognition (Do and Artières, 2009; Vinel et al., 2011). A review of discriminative methods for HMMs may be found in Nopsuwanchai (2005) and Yu and Deng (2007). In the last five years, researchers have also investigated the use of purely discriminative models for sequence recognition. Conditional Random Fields and their extension for dealing with hidden states, namely Hidden CRFs, are such models (Lafferty et al., 2001; Quattoni et al., 2007). They have already been applied to signals such as gestures and images (Quattoni et al., 2007), handwriting (Vinel et al., 2011; Do and Artières, 2006), speech (Sung and Jurafsky, 2009; Gunawardana et al., 2005; Reiter et al., 2007), eye's movements (Do and Artières, 2005).

All above discriminative approaches, discriminatively trained HMMs and HCRFs, have been shown to significantly outperform non discriminatively trained HMMs (Cheng et al., 2009; McDermott et al., 2009; Vinel et al., 2011; Sung and Jurafsky, 2009; Gunawardana et al., 2005; Sha , 2006; Fujii et al., 2011; Mahajan et al., 2004). Also, there seems to be a slight advantage to MPE and MCE among discriminative learning criterion for HMMs. Indeed, Sha and Saul (2007) and Sung and Jurafsky (2009) find that MCE and MPE slightly outperform MMI (or Conditional Maximum Likelihood, a close variant) while McDermott et al. (2009) and Cheng et al. (2009) report similar results for MCE, MPE and MMI. Next, the large margin approach is most often reported as outperforming MMI and MPE (Cheng et al., 2009; Do and Artières, 2009;

* Corresponding author.
E-mail addresses: mathieu.radenen@lip6.fr (M. Radenen), thierry.artieres@lip6.fr (T. Artières).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.08.015

Vinel et al., 2011; Sha , 2006). HCRFs are also reported as outperforming other discriminative criterion for learning HMMs (MMI, MPE, MCE) (Vinel et al., 2011; Sung and Jurafsky, 2009; Gunawardana et al., 2005; Fujii et al., 2011; Mahajan et al., 2004). Finally HCRF and large margin learning of HMMs seem to yield similar results (Vinel et al., 2011). Both methods are state of the art methods today.

One topic we are concerned with in this study is how to handle variabilty. In HMMs (as well as in HCRFs somehow) states are mutually exclusive so that it requires K states to get K different output distributions. The most popular approach to handle variability consists in increasing the number of states, in increasing the size of Gaussian mixtures in HMMs, in using context dependent unit (e.g. phone) models. These ideas are easy to implement but this quickly leads to too numerous parameters yielding over-fitting. To overcome this difficulty the speech recognition community has focused on different ways to tie parameters. Parameters can be shared between states which are acoustically indistinguishable (Hwang and Huang, 1993; Young and Woodland, 1994; Rabiner and Juang, 1993). Another strategy is to tie parameters at the distribution level (Bellegarda and Nahamoo, 1990; Paul, 1991). A pool of Gaussian is shared inside a model (partially tied), or across all models (fully tied). Yet these strategies allow capturing a local variability only, while keeping the number of parameters limited.

Our starting point is an alternative approach for handling variability. We assume that an important part of the variability between observation sequences may be modeled by a few contextual variables (which may be hidden or observed) that remain fixed all along a sequence or that vary slowly with time. For instance a sentence may be uttered quite differently according to the speaker emotion. A gesture may have more amplitude if it is performed slower, and its overall shape depends on the weight and on the height of the performer. Such a variability cannot always be removed through preprocessing or normalization and would not be captured accurately by the classical approaches above. Yet such a variability would benefit from a specific handling in HMMs and HCRFs.

Few researchers have tackled this problem by designing a HMM whose probability distribution depends on contextual variables (i.e. the context, that we note $\boldsymbol{\theta}$). Wilson and Bobick (1999) proposed Parametric hidden Markov models where the means of Gaussian distribution vary linearly as a function of the context. As the output distribution depends not only on the state but also on the context, a model may express many distribution with a limited number of additional parameters. Yu et al. (2009), Cui and Gong (2007) and Fujinaga et al. (2001) investigated rather similar approaches.

All these approaches differ by the nature of the dependency of HMM parameters to context variables, the ability to deal with dynamic context variables, i.e. evolving with time, the ability to infer context variables at test time.

We build here upon these pioneer works and propose contextual extension of both HMMs and HCRFs. We first extend parametric HMMs of Wilson and Bobick (1999) and we propose Contextual hidden Markov models (CHMMs) that rely on the parameterization of the probability distribution of a HMM (i.e. means and covariance matrices instead of means only in Wilson and Bobick (1999)) by a set of contextual variables that may vary in time. In addition, building on Gunawardana et al. (2005) which showed that a HCRF may be initialized from a trained HMM system, we provide an efficient and accurate method for learning HCRFs that exploit contextual information, that we name Contextual HCRFs (CHCRFs).

We first investigate in deep how Contextual HMMs behave with respect to standard MLE trained HMMs. Although MLE trained HMMs are clearly not the state of the art for sequence recognition and labeling, this study has some interest for at least two reasons. First CHRCF learning relies on CHMM learning. Also MLE trained HMMs are useful for modeling and synthesis tasks, handling variability in such models is then a key issue (e.g. Yu et al., 2013).

Next we focus on designing contextual HCRF that improve over HCRF which is state of the art method for signal classification and signal labeling.

In the following sections, we first motivate and introduce our modeling framework for generative models and detail the definition and the learning of our Contextual HMMs. Then we extend this framework to the definition of their discriminative counterpart, Contextual HCRFs. Finally we compare our approaches with standards HMMs and HCRFs systems on a isolated handwritten character recognition task.

2. Contextual hidden Markov model (CHMM)

In the following we focus first in Section 2.1 on the case of single Gaussian CHMM when $\boldsymbol{\theta}$ is static and remain fixed all along a sequence. Then we discuss in Section 2.2 two variants, dealing with dynamic $\boldsymbol{\theta}$ and using Gaussian mixtures.

2.1. Single Gaussian Contextual HMM with static context

2.1.1. Framework

Assume that for any observation sequence $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, where $\boldsymbol{x}_t$'s are d-dimensional feature vectors[1] we are given a set of contextual variables $\boldsymbol{\theta}$ which is a vector of dimension c. $\boldsymbol{\theta}$ might be the age and gender of a speaker for speech signals, or a set of physiological features such as height and weight for gestures, or some quantities that are computed from the input sequence $\mathbf{x}$ such as its length. We consider HMMs where means and covariance matrices depend on the available contextual information $\boldsymbol{\theta}$.

We first define the mean $\hat{\boldsymbol{\mu}}_j$ (d-dimensional vector) of the Gaussian distribution in state j to be a linear function of $\boldsymbol{\theta}$. In order to keep notations compact we consider an augmented $\boldsymbol{\theta}$ vector with all contextual variables plus a last additional component equal to 1. Hence, from now on, $\boldsymbol{\theta}$ is a c-dimensional vector $\boldsymbol{\theta} = [\text{Contextual variables}, 1]^T$ with $(c-1)$ contextual variables and a cth component equal to 1. We consider that the mean in state j is defined as:

$$\hat{\boldsymbol{\mu}}^j(\boldsymbol{\theta}) = Y^j\boldsymbol{\theta} \tag{1}$$

where Y^j is a $d \times c$ matrix. The above formulation is equivalent to writing $\hat{\mu}^j(\boldsymbol{\theta}) = V^j\boldsymbol{\theta} + \bar{\boldsymbol{\mu}}^j$ with $V^j, \bar{\boldsymbol{\mu}}^j$ and Y^j being related by: $Y^j = \left[V^j \ \bar{\boldsymbol{\mu}}_j\right]$. The vector $\bar{\boldsymbol{\mu}}_j$ is an offset vector which may be viewed as an average mean vector (eventually obtained from a traditionally learned HMM) that is modified by the linear transform part. Such a modeling has already been used in Wilson and Bobick (1999).

We go further by parameterizing covariance matrices as well. While some authors have proposed to define similarly diagonal covariance matrix that depends on external variables (Yu et al., 2009) we propose a full covariance parameterization scheme. Actually we want the covariance matrix C to be modified in such a way that each of its component $C_{u,v}$ is transformed into $C_{u,v} \times \alpha_u \times \alpha_v$ where α values depend on contextual variables $\boldsymbol{\theta}$. This allows providing an additional but limited degree of freedom to the model, allowing more expressive power while limiting overtraining risk. This may be done according to:

[1] In the following, a matrix is named with an uppercase character A, a sequence of vectors is noted in bold face $\mathbf{x}$ as well as a feature vector $\mathbf{x}$, scalars are noted in normal font.