Pattern Recognition Letters 37 (2014) 161–171

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

# Joint semi-supervised learning of Hidden Conditional Random Fields and Hidden Markov Models

Yann Soullard *, Martin Saveski, Thierry Artières

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

ARTICLE INFO

*Article history:*
Available online 6 April 2013

*Keywords:*
Hidden Markov Models
Hidden Conditional Random Fields
Semi-supervised learning
Co-training

ABSTRACT

Although semi-supervised learning has generated great interest for designing classifiers on static patterns, there has been comparatively fewer works on semi-supervised learning for structured outputs and in particular for sequences. We investigate semi-supervised approaches for learning hidden state conditional random fields for sequence classification. We propose a new approach that iteratively learns a pair of discriminative-generative models, namely Hidden Markov Models (HMMs) and Hidden Conditional Random Fields (HCRFs). Our method builds on simple strategies for semi-supervised learning of HMMs and on strategies for initializing HCRFs from HMMs. We investigate the behavior of the method on artificial data and provide experimental results for two real problems, handwritten character recognition and financial chart pattern recognition. We compare our approach with state of the art semi-supervised methods.



## 1. Introduction

Sequence classification and sequence labeling are fundamental tasks occurring in many application domains, such as speech recognition, mining financial time series, and handwriting recognition. Hidden Markov Models (HMMs) are the most popular method for dealing with sequential data (Rabiner, 1989). HMMs benefit from efficient algorithms both for inference and for training but suffer some severe drawbacks. In particular, they are traditionally learned via maximum likelihood estimation, which is a non discriminative training criterion. Many attempts have been made to overcome this limitation, relying on the optimization of a discriminant criterion like minimum error classification (Juang and Katagiri, 1992), perceptron loss (Collins, 2002), maximum mutual information (Woodland and Povey, 2002), or margin-based criterion (Sha and Saul, 2007; Do and Artières, 2009). A more recent alternative consists in defining a model of the posterior conditional probability (i.e. the probability of the labeling given the observation sequence). Hidden Conditional Random Fields (HCRFs) are such models (Quattoni et al., 2007). They are a variant of Conditional Random Fields (CRFs) (Lafferty et al., 2001) that make use of hidden states to account for the underlying structure of the data (alike in HMMs). They have been used for various signal labeling tasks, in particular for speech signals (Gunawardana et al., 2005; Reiter et al., 2007), eye movements (Do and Artières, 2005), handwriting (Do and Artières, 2006; Vinel et a., 2011), gestures and images (Morency et al., 2007) and financial time series (Soullard and Artieres, 2011).

Whatever the model one chooses to design a classification system, one needs first to gather, then to label, a sufficiently large training corpus. This often has a cost that may make the design of a good system problematic. This has motivated the study of semi-supervised learning (SSL). In SSL, classifiers are trained on both labeled samples (usually few) and unlabeled samples (usually many). A number of SSL methods have been proposed, such as entropy based methods (Grandvalet and Bengio, 2005), margin based methods (Wang et al., 2009), co-training algorithms (Blum and Mitchell, 1998) (see Mann and McCallum, 2010 for a review).

However, up to now only a few works have investigated semi-supervised learning for structured data and for sequences in particular, as we are interested in here. Some studies have investigated semi-supervised learning of HMMs for speech recognition and for text classification (Nigam et al., 2000; Inoue and Ueda, 2003; Haffari and Sarkar, 2008), but the conclusions of these works are rather limited since SSL has been shown to eventually degrade performances of supervised training (Cozman and Cohen, 2002; Mérialdo, 1994). Moreover, alternative works have focused on learning CRFs in a semi-supervised setting for language processing and biological problems, yielding some significant improvements (Jiao, 2006; Sokolovska, 2011). It is worth noting that a few of these works rely on designing a hybrid model, mixing HMMs and CRFs, where HMMs only are learned in a semi-supervised way, indirectly making the learning of CRFs semi-supervised (Sokolovska, 2011). Finally, we are not aware of any work today on SSL algorithms for complex discriminative models such as HCRFs.

* Corresponding author. Tel.: +33 1 44 27 74 91; fax: +33 1 44 27 70 00.
*E-mail addresses:* Yann.Soullard@lip6.fr (Y. Soullard), Martin.Saveski@lip6.fr (M. Saveski), 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.2013.03.028

Here we focus on semi-supervised learning for sequence classification where one wants to assign a single label to an input sequence. Extension to sequence labeling is out of the scope of the paper but should follow naturally. We propose a new algorithm for semi-supervised learning of HCRFs for sequence classification. It relies on an iterative joint learning of a pair of generative and discriminative models, namely HMMs and HCRFs. This paper is an extension of our previous work in Soullard and Artieres (2011), and improves on it in several ways. First, we describe in more detail our approach, in particular the initialization scheme of HCRF from Full Covariance matrix Gaussian HMMs. Second, we propose and investigate a few variants of our method. Third, we provide new results on artificial data for an improved understanding of the behavior of the method. Fourth, we provide additional results on real datasets and provide a thorough experimental comparison of our approach with state of the art SSL methods that were already proposed for CRFs and that we extended to HCRFs.

We first present related works on semi-supervised learning in Section 2, then we detail in Section 3 our strategy for initializing HCRFs from Full Covariance matrix Gaussian HMMs. Next, we discuss the motivation of our approach, which we present in detail in Section 4. We report experimental results on artificial data in Section 5 and we investigate in Section 6 the behavior of our approach for two real problems, handwritten character recognition and financial chart pattern classification.

## 2. State of the art in semi-supervised learning

Here, we review the main semi-supervised learning approaches (Zhu and Goldberg, 2009), with a particular focus on methods that have been used or that could be extended for learning markovian models such as HMMs and CRFs.

In this study, we focus on classification where training samples are couples $(\mathbf{x}, y), \mathbf{x} \in \mathcal{X}$ is an input sample (e.g. a sequence) and where $y \in \mathcal{Y}$ is its class (i.e. label).[1] We denote $L = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^{|L|}, y^{|L|})\}$ as the set of labeled training samples, with $|L|$ being its cardinal, and $U = \{\mathbf{x}^{|L|+1}, \ldots, \mathbf{x}^{|L|+|U|}\}$ stands for the set of unlabeled training samples. Also, in the following we will systematically use $\Theta$ to denote the set of parameters of generative models (e.g. HMMs) and $\Lambda$ to denote the set of parameters of discriminative models (e.g. CRFs).

### 2.1. Mixture approach

The *mixture approach* consists of learning a mixture of generative models, one for each class, through an Expectation Maximization (EM) like algorithm. In Nigam et al. (2000), the EM algorithm was applied on a mixture of multinomial distributions for text classification while in Baluja (1998) it was applied on a face orientation discrimination task. This approach has been applied to HMMs in Nigam et al. (2000); Inoue and Ueda (2003). The objective criterion to be maximized is defined as:

$$\mathcal{L}(\Theta) = \frac{(1-\gamma)}{|L|} \sum_{i=1}^{|L|} log\, p(\mathbf{x}^{(i)}, y^{(i)}|\Theta) + \frac{\gamma}{|U|} \sum_{j=|L|+1}^{|L|+|U|} log\, p(\mathbf{x}^{(j)}|\Theta) \quad (1)$$

where $\gamma \in [0,1]$ is a parameter that allows tuning of the relative influence of labeled data and unlabeled data. The fully supervised and the fully unsupervised cases are specific instances when $\gamma$ is respectively set to 0 and to 1 (Ji et al., 2009). Although it is a simple and attractive idea, this approach may degrade HMMs' performances (Cozman and Cohen, 2002; Mérialdo, 1994), especially if the number of labeled samples is too small.

[1] Note that we use bold font to denote sequences, e.g. $\mathbf{x}$, while we use normal font for static patterns, vector or scalar, e.g. $y$.

### 2.2. Minimum entropy

*Minimum entropy* regularization is a popular technique (Grandvalet and Bengio, 2005). It aims at reducing uncertainty on the labeling of unlabeled samples. the method is extended in Jiao (2006) to the learning of CRF and is used with the following regularized objective function:

$$\mathcal{L}_\gamma(\Lambda) = -\frac{||\Lambda||^2}{2} + \sum_{i=1}^{|L|} log\, p(y^{(i)}|\mathbf{x}^{(i)}, \Lambda) + \gamma \sum_{j=|L|+1}^{|L|+|U|} \sum_{y \in \mathcal{Y}} p(y|\mathbf{x}^{(j)}, \Lambda) log\, p(y|\mathbf{x}^{(j)}, \Lambda) \quad (2)$$

The above objective combines conditional entropy for unlabeled samples and conditional likelihood for labeled samples. A similar approach was taken in Wang et al. (2009) by defining the objective function as the combination of the conditional likelihood of the labeled data and of the mutual information for the unlabeled data.

### 2.3. Co-training

*Co-training* has been popularized by Blum and Mitchell (1998) for static patterns. It assumes that the features used to represent a sample may be split into two sets of features, or views, (every sample then has two representations, one for each view) and that these two views are sufficient for a correct classification. Learning consists of first training two classifiers, one for each view. Then one selects the unlabeled samples for which one classifier is most confident and puts these samples together with the classifier's predictions into the training set of the other classifier. This process is repeated iteratively. The approach is extended in Wang and Zhou (2007) to the case where two classifiers are trained on the same view and showed that co-training may work well provided the classifiers are different enough.

Co-training has also been investigated with some success for learning generative markovian models. In particular, the standard co-training algorithm was applied in Khine et al. (2008) to HMMs for singing voice detection and co-training of HMMs and of neural networks was experimented in Frinken et al. (2009) for handwriting recognition.

### 2.4. Hybrid methods

A few methods have been proposed to mix generative and discriminative methods (Bishop and Lasserre, 2007; Bouchard, 2007 ). These methods rely on the idea that semi-supervised learning is more natural for learning generative models with a non discriminative criterion through, e.g. the *mixture approach*. In Bouchard (2007), the parameters of generative models are learnt by optimizing a combination of a non discriminative criterion (e.g. likelihood) and of a discriminative criterion (conditional likelihood), where the non discriminative criterion is computed for all training data (labeled and unlabeled) while the discriminative criterion concerns labeled training data only. Furthermore, some authors proposed in Bishop and Lasserre (2007) to learn two linked sets of parameters of generative models, one parameter set with the non discriminative criterion (on the entire training dataset) and the other parameter set with the discriminative criterion (on the labeled training dataset) with the following objective function:

$$\mathcal{L}(\Theta, \Lambda) = \sum_{i=1}^{|L|} \log p(y^{(i)}|\mathbf{x}^{(i)}, \Lambda) + \sum_{j=1}^{|L|+|U|} \log p(\mathbf{x}^{(j)}|\Theta) + log(p(\Theta, \Lambda)) \quad (3)$$

where $p(\Theta, \Lambda)$ is a prior that links the two parameter sets. It allows blending generative and discriminative approaches. If the prior is uniform, the generative and discriminative models are