Pattern Recognition 46 (2013) 1569–1578

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A conditional random field-based model for joint sequence segmentation and classification

Sotirios P. Chatzis [a,*], Dimitrios I. Kosmopoulos [b,c], Paul Doliotis [c,d]

[a] Department of Electrical Engineering, Computer Engineering, and Informatics, Cyprus University of Technology, Cyprus
[b] Department of Computer Science, Rutgers University, 08854 NJ, USA
[c] NCSR Demokritos, Institute of Informatics and Telecommunications, GR 15310, Greece
[d] University of Texas at Arlington, Computer Science and Engineering, 76013 TX, USA

ARTICLE INFO

Article history:
Received 5 March 2012
Received in revised form
15 September 2012
Accepted 29 November 2012
Available online 20 December 2012

Keywords:
Conditional random field
Sequence segmentation
Sequence classification

ABSTRACT

In this paper, we consider the problem of joint segmentation and classification of sequences in the framework of conditional random field (CRF) models. To effect this goal, we introduce a novel dual-functionality CRF model: on the first level, the proposed model conducts sequence segmentation, whereas, on the second level, the whole observed sequences are classified into one of the available learned classes. These two procedures are conducted in a joint, synergetic fashion, thus optimally exploiting the information contained in the used model training sequences. Model training is conducted by means of an efficient likelihood maximization algorithm, and inference is based on the familiar Viterbi algorithm. We evaluate the efficacy of our approach considering a real-world application, and we compare its performance to popular alternatives.

1. Introduction

The problem of predicting from a set of observations, a set of corresponding labels, that are statistically correlated within some combinatorial structures like chains or lattices is of great importance, because it appears in a broad spectrum of application domains including annotating natural language sentences (e.g., parsing, chunking, named entity recognition), labeling biological sequences (e.g., protein secondary structure prediction), and classifying regions of images (e.g., image segmentation with object recognition), to name just a few.

Graphical models are a natural formalism for exploiting the dependence structure among entities. Traditionally, graphical models have been used to represent the joint probability distribution $p(\boldsymbol{y},\boldsymbol{x})$, where the variables $\boldsymbol{y}$ represent the attributes of the entities that we wish to predict, and the variables $\boldsymbol{x}$ represent our observed knowledge about the entities. But modeling the joint distribution can lead to difficulties, because it requires modeling the distribution $p(\boldsymbol{x})$, which can include complex dependencies. Modeling these dependencies among inputs can lead to intractable models, but ignoring them can lead to reduced performance. A solution to this problem is to directly model the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$, which is sufficient for classification. Indeed, this is the approach taken by conditional random fields (CRFs) [1].

A conditional random field is simply a log-linear model representing the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ with an associated graphical structure. Because the model is conditional, dependencies among the observed variables $\boldsymbol{x}$ do not need to be explicitly represented, affording the use of rich, global features of the input. For example, in natural language tasks, useful features include neighboring words and word bigrams, prefixes and suffixes, capitalization, membership in domain-specific lexicons, and semantic information from sources such as WordNet [2]. During the last years, we have witnessed an explosion of interest in CRFs, as they have managed to achieve superb prediction performance in a variety of scenarios, thus being one of the most successful approaches to the structured output prediction problem, with successful applications including text processing, bioinformatics, natural language processing, and computer vision [1,3–8].

In this paper, we focus on linear-chain CRFs. Linear-chain CRFs, the basic probabilistic principle of which is illustrated in Fig. 1(a), are conditional probability distributions over label sequences, which are conditioned on the observed sequences [1,2]. Hence, in conventional linear-chain CRF formulations, an one-dimensional first-order Markov chain is assumed to represent the dependencies between the modeled data. In our work, we seek to provide a novel CRF-based model for joint segmentation and classification of observed sequences. Indeed, joint sequence segmentation and classification is a perennial problem that occurs in several application areas, such as object and behavior recognition in computer

* Corresponding author. Tel.: +35 725002041.
E-mail addresses: soteri0s@me.com (S.P. Chatzis), dkosmo@ieee.org (D.I. Kosmopoulos), doliotis@iit.demokritos.gr (P. Doliotis).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.11.028

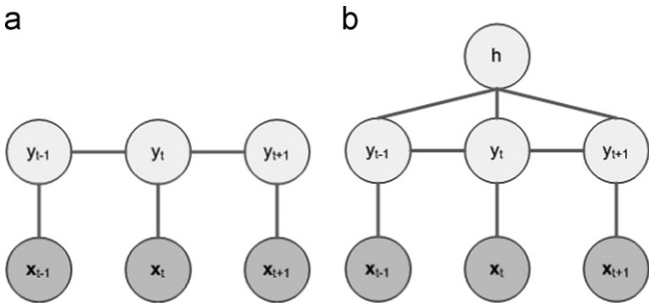

Fig. 1. Linear-chain conditional random fields: the conventional and the proposed approach. The brightly colored nodes denote a random variable y_t, and the shaded nodes $\mathbf{x}_t$ have been set to the respective observed values. h denotes the sequence label. (a) Conventional CRF. (b) Proposed CRF.

vision applications (e.g., [9,10]), speech analysis (e.g., [11]), and bioinformatics [12]. By jointly treating the two tasks of sequence decoding (segmentation) and classification, one can more effectively exploit the available information, thus allowing for a potentially considerable increase in the obtained algorithm performance [10].

Towards this end, in this paper we propose a novel dual-functionality CRF (DF-CRF) model for joint sequence segmentation and classification. Our proposed model comprises two levels of functionality. In the first level, the observed sequences are segmented, using a variant of the familiar Viterbi algorithm. In the second level, classification of the observed sequences is performed. Both these procedures are conducted concurrently and in a synergetic fashion, thus optimally exploiting the information acquired from the available training data. Model training is effected by means of a computationally efficient likelihood maximization algorithm. We evaluate our novel approach in a real-world visual workflow segmentation and recognition application; as we show, our proposed approach offers considerable improvement over hidden Markov models (HMMs) [13], standard CRFs, as well as hidden conditional random field (HCRF) models [9], a method related to the DF-CRF, but considering that the sequence segment labels are not observable and, hence, comprise a latent variable of the model.

The remainder of this paper is organized as follows: in Section 2, a brief introduction to CRFs is provided. In Section 3, the proposed DF-CRF model is introduced, its inference algorithms are derived, and we discuss the differences between our proposed approach and HCRFs. In Section 4, we apply our model to a real-world application dealing with visual workflow recognition and decoding, using challenging datasets obtained from the assembly lines of an automobile manufacturer. We compare our method's performance to HMMs, standard CRFs, and HCRF models. Finally, in the concluding section of this paper, we summarize our contribution and results.

2. Conditional random fields

In the following, we provide a brief introduction to linear-chain CRF models, which constitute the main research theme of this paper. For a more detailed account of CRF models, the interested reader may refer to [2].

Linear-chain CRFs typically assume dependencies encoded in a left-to-right chain structure. Formally, linear-chain CRFs are defined in the following fashion: Let $\{\boldsymbol{x}_t\}_{t=1}^{T_X}$ be a sequence of observable random vectors, and $\{\boldsymbol{y}_t\}_{t=1}^{T_Y}$ be a sequence of random vectors that we wish to predict. Typically, the model is simplified by assuming that the lengths of the two sequences are equal, i.e., $T_X = T_Y = T$, and that the predictable variables are scalars defined on a vocabulary comprising K words, i.e., $y_t \in \mathcal{Y}$, with $\mathcal{Y} = \{1,\ldots,K\}$, whereas the observable variables are usually defined on a high-dimensional real space, $\boldsymbol{x}_t \in \mathcal{X}$, with $\mathcal{X} \subseteq \mathbb{R}^{\zeta}$. Then, introducing the notation $\boldsymbol{x} = ([\boldsymbol{x}_t^T]_{t=1}^T)^T$, and $\boldsymbol{y} = [y_t]_{t=1}^T$, a first-order linear-chain CRF defines the conditional probability for a label sequence $\boldsymbol{y}$ to be given by

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left[\sum_{t=2}^{T} \phi_t(y_t, y_{t-1}, \boldsymbol{x}_t) + \phi_1(y_1, \boldsymbol{x}_1)\right], \tag{1}$$

where $\phi_t(\cdot)$ is the local *potential* (or *score*) *function* of the model at time t, and $Z(\boldsymbol{x})$ is a partition function that ensures the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ of a state sequence $\boldsymbol{y}$ will always sum to one

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left[\sum_{t=2}^{T} \phi_t(y_t, y_{t-1}, \boldsymbol{x}_t) + \phi_1(y_1, \boldsymbol{x}_1)\right]. \tag{2}$$

In this work, we will be assuming that the potential functions of the postulated linear-chain CRFs can be written in the form

$$\phi_t(y_t, y_{t-1}, \boldsymbol{x}_t) = \phi_t^1(y_t, \boldsymbol{x}_t) + \phi_t^2(y_t, y_{t-1}), \tag{3}$$

$$\phi_1(y_1, \boldsymbol{x}_1) = \phi_1^1(y_1, \boldsymbol{x}_1) + \phi_1^2(y_1), \tag{4}$$

where the $\phi_t^1(y_t, \boldsymbol{x}_t)$ and the $\phi_t^2(y_t, y_{t-1})$ are the *unary* and *transition potentials* of the model, respectively, centered at the current time point. Note that, in the above definition, we have considered that the transition potentials $\phi_t^2(y_t, y_{t-1})$ do not depend on the observations $\boldsymbol{x}_t$, but, instead, a given transition, say from state i to state j, always receives the same transition potential function value regardless of the input. Such a model formulation is usually referred to as a hidden Markov model (HMM)-like linear-chain CRF [2]. We will be considering this form of transition potentials throughout this work; however, our results can be easily extended to any other formulation, where the transition potentials are assumed to also depend on the observed input variables $\boldsymbol{x}$.

Regarding the form of the unary and transition potentials usually selected in the literature, the most typical selection consists in setting

$$\phi_t^1(y_t, \boldsymbol{x}_t) = \sum_{i=1}^{K} \delta(y_t - i)\boldsymbol{\omega}_i^T \boldsymbol{x}_t \tag{5}$$

and

$$\phi_t^2(y_t, y_{t-1}) = \sum_{i=1}^{K}\sum_{j=1}^{K} \delta(y_t - j)\delta(y_{t-1} - i)\xi_{ij} \tag{6}$$

with

$$\phi_1^2(y_1) = \sum_{i=1}^{K} \delta(y_t - i)\xi_i, \tag{7}$$

where $\delta(\sigma)$ is the Dirac delta function, the parameters $\boldsymbol{\omega}_i$ are the prior weights of an observation emitted from state i, the parameters ξ_{ij} are related to the prior probabilities of the transition from state i to state j, and the parameters ξ_i are related to the prior probabilities of being at state i at the initial time point $t = 1$. Estimates of these parameters are obtained by means of model training, which consists in maximization of the log of the model likelihood, given by (1). For this purpose, usually quasi-Newton optimization methodologies are employed, such as the BFGS algorithm [14], or its limited memory variant (L-BFGS) [15], which, indeed, is the most commonly used method in the CRF literature [1,2].

Note that computation of the model likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ entails calculation of the sum $Z(\boldsymbol{x})$ defined in (2). This can be effected in a computationally efficient manner using the familiar forward–backward algorithm [16,13], widely known from the HMM literature.