Neurocomputing 150 (2015) 449–457

Contents lists available at ScienceDirect

Neurocomputing

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

Greedy ensemble learning of structured predictors for sequence tagging

Minyoung Kim *

Department of Electronics & IT Media Engineering, Seoul National University of Science & Technology, Seoul 139-743, South Korea

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form
23 April 2014
Accepted 2 May 2014
Available online 16 October 2014

Keywords:
Structured output classification
Functional gradient boosting
Conditional random fields
Sequence tagging

ABSTRACT

We tackle the sequence tagging problem where the multiple output labels to be predicted are correlated with one another in a complex manner. Due to the ability of capturing statistical dependency in the output variables, the structured models like conditional random fields (CRFs) have received significant attention recently. For computational issues, however, the CRF typically assumes rather simple restricted dependency structures like chains, which can limit its prediction performance considerably when the true data generation processes do not match with the model assumption. In this paper we propose novel algorithms to learn an ensemble of predictor models to boost the overall prediction accuracy. By looking at the frame-wise predictor inferred from a structured CRF model as a weak classifier, the ensemble learning can be formulated within the (functional gradient) boosting framework. Similar to the conventional single-output boosting algorithms, our methods produce the frame-wise importance weights on training data at each stage that gives crucial guidance of which output variables to focus on more (and which less) in learning the next-stage CRF model. The stage-wise learning reduces to the weighted frame-wise conditional likelihood maximization, which can be done as fast as the conventional CRF learning. Our approaches differ from the ordinary single-output boosting in that the base predictors are not learned independently across different frames, yet they are derived from the same structural CRF model, hence tightly clamped with each other to impose overall smoothness and consistency constraints. We demonstrate the improved prediction accuracy on several sequence tagging problems.

1. Introduction

In this paper we deal with the sequence tagging problem aimed for accurate prediction of the multiple output class labels that are correlated with one another in a complex manner. Typically, the output variables exhibit temporal dependencies, where the related important application problems include motion sequence segmentation, theme-based video annotation, facial emotion estimation in video, part of speech tagging or parsing of natural language sentences, and protein secondary structure prediction, to name just a few.

To capture the complex statistical correlation between the output variables, the discriminative structured models like conditional random field (CRF) emerged in the last decade, and have become the major computational tools for accurate structured output prediction [1–7]. The predictive power of the CRF originates from its ability to capture the statistical dependency of output variables via conditional probabilistic modeling. Although there exist other probabilistic approaches like the popular hidden Markov model (HMM), the HMM is a generative model that represents the joint distribution of input and output. If the predictive performance is the main concern, the conditional models like CRFs can circumvent the difficult modeling effort for the input distribution, and focus solely on capturing the impacts of inputs on outputs, resulting in higher prediction accuracies than generative models [8,9,1,3].

Even though it has achieved great success in many related fields, the CRF model typically assumes, mainly due to computational reasons, rather simple restricted dependency structures like chains, which can limit its prediction performance considerably especially when the true data generation processes do not match with the underlying model assumption. While there have been other attempts to enlarge the model representational capacity by incorporating latent variables into the CRF [4,10,11], they introduce additional computational complexity in the probabilistic inference, also suffering from local optima issues.

In this paper we consider to build more accurate structured output classifiers without introducing further computational overhead. The main idea is motivated from the general ensemble/boosting paradigm: the weak predictors that are inferred from the CRF with a simple chain structure can be combined together in a

* Tel.: +82 2 9709020; fax: +82 2 9707903.
E-mail address: mikim21@gmail.com

http://dx.doi.org/10.1016/j.neucom.2014.05.079
0925-2312/

principled manner to lead to a much stronger classifier. In our case, the frame-wise classifier, which predicts each output variable y_t at each frame t in the sequence, is considered to be augmented/boosted to yield a powerful ensemble prediction model. Note however that the frame-wise predictors across different frames are all clamped to one other within the same structured model (CRF) to impose the smoothness and consistency constraints on the entire output labels.

How to form an ensemble model from the base predictors is the main theme of the paper. We propose two greedy algorithms based on the functional gradient boosting framework. The first strategy is to build a probabilistic mixture model for the frame-wise output distributions. We provide a stage-wise predictor selection scheme that maximizes the conditional class likelihood objective functional in a greedy fashion. The second is to adopt the logit classifier as a base predictor, where we minimize the exponential loss (a smooth approximation of the 0/1 loss) incurred by the addition of the base predictor. In both strategies, we use the steepest gradient search in function space to derive how a newly added predictor can be learned.

Similar to the conventional single-output boosting algorithms (e.g., [12]), the proposed methods produce the frame-wise importance weights on training data at each stage, which gives crucial guidance of which output frames to focus on more (and which less) in learning the next-stage CRF model. Interestingly, the proposed data weighting schemes are intuitively appealing by assigning higher weights on the misclassified frames, and vice versa. In computational aspects, the stage-wise learning reduces to the weighted frame-wise posterior likelihood maximization, which can be done as fast as the conventional CRF learning.

It should be noted that our approaches differ from the ordinary single-output boosting in that the base predictors are not learned independently across different frames, yet they are derived from the same structural CRF model. Hence the frame-wise predictors at each stage are coupled with one another within the same CRF, which can yield predicted results that are smooth over the sequence in a temporal sense. Thus, unlike building a set of independent predictors one for each of multiple outputs, we do consider the overall performance on the structured data; not merely focusing on improving the accuracy at a single frame, but for *all* outputs.

The paper is organized as follows: After briefly describing the formal problem setup and introducing notations, we give some background on CRF with several recent CRF learning methods in Section 2. The proposed ensemble learning algorithms are discussed in Section 3 while in the Appendix, the detailed derivations for the weighted frame-wise conditional likelihood maximization estimator used for learning stage-wise CRFs are described. In Section 4 we demonstrate the improved prediction accuracy achieved by two ensemble approaches on several important sequence tagging problems.

1.1. Problem setup and notations

We denote the output random variables in sequence by bold-faced $\mathbf{Y}$, comprised individual variables y_t at frames $t=1,\ldots,T$ (i.e., $\mathbf{Y}=y_1\cdots y_T$). Each output variable is discrete-valued, taking one of the K different class labels, that is, $y_t\in\{1,\ldots,K\}$. The predictor variables or the observation features are denoted by $\mathbf{X}$ that admit a similar sequence structure as the output $\mathbf{Y}$. Specifically $\mathbf{X}=\mathbf{x}_1\cdots\mathbf{x}_T$, and the feature vector at each frame $\mathbf{x}_t$ is assumed to be a p-variate vector. Notice that we do not assume that the sequence length T is fixed, but can vary from instance to instance.

We exemplify two popular applications of the sequence tagging, the automatic speech recognition and the facial emotion tagging in video. In the former, one aims to predict the transcript sequence $\mathbf{Y}=y_1\ldots y_T$ from the input sequence, $\mathbf{X}=\mathbf{x}_1\ldots\mathbf{x}_T$ comprised certain features extracted from speech signals. The predicted tag y_t indicates the uttered word (out of K words) that corresponds to the speech feature $\mathbf{x}_t$ at time t. In the latter, given a sequence of input features $\mathbf{X}$ (e.g., $\mathbf{x}_t$ contains certain image features extracted from the t-th video frame), the goal is to accurately predict the emotion intensity level (say, $y_t\in\{\texttt{Neutral},\texttt{Increasing},\texttt{Apex}\}$) for each frame t.

In this paper we consider the sequence tagging problem within the supervised learning setup: one is given a dyadic set of training data, $\mathcal{D}=\{(\mathbf{X}^i,\mathbf{Y}^i)\}_{i=1}^n$, n i.i.d. samples from an underlying but unknown distribution $P(\mathbf{X},\mathbf{Y})$. Our goal is to learn an accurate structured prediction function $\mathbf{Y}=h(\mathbf{X})$. Among several machine learning approaches, the conditional random field (CRF) aims to represent the probabilistic conditional distribution $P(\mathbf{Y}|\mathbf{X})$ to account for the complex statistical dependency among the input/output variables. Then the predictor can be determined from the model by statistical inference, $h(\mathbf{X})=\arg\max_{\mathbf{Y}}P(\mathbf{Y}|\mathbf{X})$. In the next section we briefly review the CRF, then our ensemble CRF learning approaches are proposed in the subsequent section.

2. Background on CRF

The conditional random field aims to represent the conditional distribution $P(\mathbf{Y}|\mathbf{X})$ as the Gibbs form:

$$P(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})=\frac{e^{s(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})}}{Z(\mathbf{X};\boldsymbol{\theta})}\quad\text{where } Z(\mathbf{X};\boldsymbol{\theta})=\sum_{\mathbf{Y}\in\mathcal{Y}}e^{s(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})}. \tag{1}$$

The denominator $Z(\mathbf{X};\boldsymbol{\theta})$ is a normalizer (often called the *partition function*) to enforce a distribution. Here we use $\mathcal{Y}$ to denote a set of all possible output sequence realizations. The relationship between input and output variables are encoded in the score function $s(\mathbf{X},\mathbf{Y})$, where one typical way to form it is via (generalized) log-linear modeling:

$$s(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})=\boldsymbol{\theta}^\top\cdot\boldsymbol{\Psi}(\mathbf{X},\mathbf{Y}). \tag{2}$$

That is, the score is defined as the inner product between the model parameters $\boldsymbol{\theta}$ and the joint feature vector $\boldsymbol{\Psi}(\mathbf{X},\mathbf{Y})$. The latter is also referred to as the *sufficient statistics* of the log-linear model. One can interpret the score as the *negative energy* that assigns higher values to more likely configurations, and vice versa.

Indeed, it is the joint feature vector $\boldsymbol{\Psi}(\mathbf{X},\mathbf{Y})$ that determines the model's dependency structure on the input/output variables. The conditional distribution $P(\mathbf{Y}|\mathbf{X})$ can be factorized into smaller terms according to the dependency structure we define in the feature function. More specifically, when we let $\mathcal{C}$ be the set of all cliques in the output graph (variables within the clique roughly indicate inter-dependencies among the variables, cf., the Hammersley–Clifford theorem [13]), one defines the score function as $s(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})=\sum_{c\in\mathcal{C}}\boldsymbol{\theta}_c^\top\cdot\boldsymbol{\Psi}_c(\mathbf{X},\mathbf{Y}_c)$ where $\mathbf{Y}_c$ indicates the output variables confined to the clique $c\in\mathcal{C}$ (similarly for $\boldsymbol{\theta}_c$ and $\boldsymbol{\Psi}_c$). This results in the factorized distribution $P(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})\propto\prod_{c\in\mathcal{C}}\exp(\boldsymbol{\theta}_c^\top\cdot\boldsymbol{\Psi}_c(\mathbf{X},\mathbf{Y}_c))$.

Accordingly, how we design the output graph structure $G=(V,E)$ (e.g., cliques), significantly affects model's representational capacity. Generally, the more complex the graph structure is, one can have the richer distribution family, however, at the expense of higher complexity for accurate inference within the model. For tractable inference, the most widely used graph structure is the simple chain, for which the cliques are confined to the two adjacent variables y_t and y_{t-1}.

For the chain-structure models, cliques are restricted to be pairwise, and we group them into two different types: *node* cliques and *edge* cliques. The node features are denoted by $\boldsymbol{\Psi}_t^{(V)}(\mathbf{X},y_t)$ and the edge features by $\boldsymbol{\Psi}_t^{(E)}(\mathbf{X},y_t,y_{t-1})$. We also split