Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised learning of local structured output predictors

Xin Du

School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Article history:
Received 16 January 2016
Received in revised form
15 February 2016
Accepted 21 February 2016

Keywords:
Machine learning
Structured output
Semi-supervised learning
Local linear regression
Gradient descent

ABSTRACT

In this paper, we study the problem of semi-supervised structured output prediction, which aims to learn predictors for structured outputs, such as sequences, tree nodes, vectors, etc., from a set of data points of both input–output pairs and single inputs without outputs. The traditional methods to solve this problem usually learn one single predictor for all the data points, and ignore the variety of the different data points. Different parts of the data set may have different local distributions and require different optimal local predictors. To overcome this disadvantage of existing methods, we propose to learn different local predictors for neighborhoods of different data points, and the missing structured outputs simultaneously. In the neighborhood of each data point, we proposed to learn a linear predictor by minimizing both the complexity of the predictor and the upper bound of the structured prediction loss. The minimization is conducted by gradient descent algorithms. Experiments over four benchmark data sets, including DDSM mammography medical images, SUN natural image data set, Cora research paper data set, and Spanish news wire article sentence data set, show the advantages of the proposed method.

1. Introduction

Machine learning refers to the problem of learning a predictive model to predict an output from an input data point [1–9]. The forms of output are various, usually including binary class label and continuous response. The problem of predicting binary class label is called classification [10–18], while the problem of predicting continuous response is called regression [19–22]. Both of these problems have many applications, such as computer vision, natural language processing, bioinformatics, and finance. However, in many of these applications, the forms of outputs of the prediction may be beyond binary class labels and continuous responses. For example, in the part-of-speech tagging problem of natural language processing, given a sequence of words, we want to predict the tags of the part-of-speech of the works, and the output of the prediction is a sequence of parts-of-speech [23–27]. In the problem of hierarchical image classification problem, the class labels of images are organized as a tree structure, and the outputs of the prediction problem are the leaves of a tree [28–31]. In this case, the predictive models designed for binary class labels and continuous responses cannot handle these output forms, and new predictive model should be developed. The output forms other than binary labels and continuous responses are called structured outputs. The structured outputs include a wide range of types of outputs, such as sequences, vectors, graph nodes, tree leaves, etc. The problem of learning predictive models to predict unknown structured outputs is called as structured output prediction [32–36]. Given a set of input–output data points, where the outputs are structured, this problem usually learns a predictive model to match the input–output relationship. Most existing methods designed to solve assume that in the training set, all the input data points have their corresponding outputs. However, in real-world application, many outputs are not available for the inputs [37–40]. The training set is composed of two parts. One part is a set of input–output pairs, which is called labeled set. The other part is a set of single input data points with missing outputs, and this set is called unlabeled data set. Learning from such a training set is called semi-supervised learning [41–43]. In this paper, we invest the problem of learning structured output predictors from such a training set. This problem is called semi-supervised structured output prediction.

1.1. Related works

Our work is a novel semi-supervised structured output prediction method, thus we introduce the related works of this direction. There are a number of existing algorithms for this problem, which are briefly introduced as follows:

1. Altun et al. [37] proposed the problem of semi-supervised learning with structured outputs. Moreover, a novel discriminative approach was also proposed to use the manifold of input

E-mail address: xindu.njtu@gmail.com

http://dx.doi.org/10.1016/j.neucom.2016.02.086
0925-2312/

features of both labeled and unlabeled data points. This approach is based on the semi-supervised maximum-margin formulation. It is an inductive algorithm, and it can be easily extended to new-coming test data points.
2. Brefeld and Scheffer [38] proposed to solve the problem of semi-supervised structured output prediction by learning in the space of input–output space, and using co-training method. This method is based on the assumption that the multiple structured output predictors should be consent with each other. Based on this assumption, the structural support vector machine is extended to the argued input–output space.
3. Suzuki et al. [39] proposed a hybrid method to solve the problem of semi-supervised structured output learning. This method combines both the generative and discriminative methods. The objective of this method is composed of log-linear forms of both discriminative structured predictor and generative model. The generative model is used to incorporate unlabeled data points. The discriminant functions are enhanced by the unlabeled data points provided by the generative model.
4. Jiang et al. [40] proposed to regularize the structured outputs by the manifold constructed from the input space directly. This method constructs a nearest neighbor graph from the input features, and use it to represent the manifold. Then the manifold is used to regularize the learning of the missing outputs of the unlabeled data points. The outputs and the predictor are learned simultaneously, and they regularize each other in the learning process.

Our work approximates the upper bound of the structured loss and is inspired by the lower bound approximation of the structure learning of the Bayesian network [13,4]. Thus we also discuss the works of bound approximation technologies of [13,4]:

1. Fan et al. [13] proposed to tighten the upper and lower bounds of the breadth-first branch and bound algorithm for the learning of Bayesian network structures. The informed variable groupings are used to create the pattern databases to tighten the lower bounds, while the anytime learning algorithm is used to tighten the upper bound. These strategies show good performance in the learning process of the Bayesian network structures. The work of [13] is a contribution of major significance to the bound approximation community, and our upper bound approximation method is also based on these strategies.
2. Fan et al. [4] further proposed to improve the lower bound function of static k-cycle conflict heuristic for the learning of Bayesian network structures. This work is used to guild the search of the most promising search spaces. It uses a partition of the random variables of a data set, and the further research is based on the importance of the partition. A new partition method was proposed, and it uses the information extracted from the potentially optimal parent sets.

1.2. Our contributions

All the mentioned semi-supervised structured output prediction methods learn one single predictor for the entire data set. However, we observe that a training set, the local distributions of neighborhoods play important roles in the problem of modeling of both input and structured outputs. It is extremely important to respect the local distributions when the structured output predictor is learned. This is even more important for learning from semi-supervised data sets. This is because for this type of data set, only a few data points have available structured outputs, and the structured outputs of all other data points are missing. To learn the missing structured outputs, we need to explore the connections between different data points, so that we may propagate the structured outputs from the labeled set to the unlabeled set. It has been shown that using local connections is an effective way to model the connections among different data points [40]. To explore the local distributions, one option is to construct a nearest neighbor graph, and use it to regularize the learning of the predictors. More specifically, with the nearest neighbor graph, we hope that the neighboring data points can obtain similar structured outputs from the predictor [44,40]. However, one single predictor is usually not enough to characterize multiple local distributions, thus even when we use the neighborhood graph to regularize the learning of the predictor, it is still not guaranteed that the local distributions are sufficiently modeled with regard to the structured output prediction problem. This is an even more serious problem when it is applied to a semi-supervised data set. With such a data set, only a few data points have corresponding structured outputs, while most of the data points do not. The learned predictor can easily fit to the labeled data points.

To solve this problem, we propose to learn multiple local linear structured output predictor for different neighborhoods to model the local distributions, instead of learning one single predictor for the entire data [45,46]. Moreover, we also propose to learn the missing structured outputs for a semi-supervised data set simultaneously. For each data point, we propose to present the local distribution around this data point by its k nearest neighborhood, and model it by learning a local linear structured output predictor. To learn the parameters of this local predictor, we propose to minimize an upper bound of the structured losses of the data points in this neighborhood, and the squared ℓ_2 norm of the predictor parameter vector. In this process, the predicted structured outputs are compared to the learned structured outputs. The learning of the structured outputs is done simultaneously by the true structured outputs of the labeled data points and the local predictors. Some data points are shared by different neighborhoods, and they play the role of bridging different local distributions to learn a complete manifold. To solve the problem, we develop an iterative algorithm by using gradient descent method.

1.3. Paper organization

The rest of this paper is organized as follows. In Section 2, we model the learning problem and present the optimization methods for the problem. In Section 3, the iterative algorithm for the learning process and the algorithm for the test process are both introduced. In Section 4, the experiments on four benchmark data sets are presented, including DDSM Mammography medical image data set, SUN natural image data set, Cora research paper data set, and Spanish news wire article sentence data set. In Section 5, the paper is concluded. In Section 6, we discuss the future works.

2. Problem modeling and optimization

2.1. Problem modeling

Suppose we have a training set of n data points, $\mathcal{X} = \mathcal{L} \cup \mathcal{U}$, which is composed of a labeled subset, $\mathcal{L}$, and an unlabeled subset, $\mathcal{U}$. $\mathcal{L}$ contains l data points of input–output pairs, $\mathcal{L} = \{(\mathbf{x}_i, \bar{y}_i)\}_{i=1}^{l}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the d-th dimensional feature vector input of the i-th data point, $\bar{y}_i \in \mathcal{Y}$ is the true structured output of the i-th data point, $\mathcal{Y}$ is the space of structured outputs. $\mathcal{U}$ contains $u = n - l$ data points of inputs without outputs, $\mathcal{U} = \{\mathbf{x}_i\}_{i=l+1}^{n}$. The structured output prediction problem is to learn predictors to predict the structured outputs from the inputs from the training set. We proposed to learn a local predictor for the neighborhood of each data point, instead of learning one single predictor for all the data points. We present the neighborhood of