Semi-Supervised Structure Learning

Yasemin Altun, David McAllester
Toyota Technological Institute at Chicago, IL
{altun,mcallester}@tti-c.org

Extended Abstract

Discriminative learning framework is one of the very successful fields of machine learning. The methods of this paradigm, such as Boosting, and Support Vector Machines have significantly advanced the state-of-the-art for classification by improving the accuracy and by increasing the applicability of machine learning methods.

Recently there has been growing interest to generalize discrimative learning methods to handle structured labels. For example labeling a word sequence with a part of speech sequence or labeling a word sequence with a parse tree. A variety of learning methods have been generalized to the structured case including logistic regression, perceptron (voted and dual), boosting, SVMs and kernel logistic regression (See [1] for a review on this line of research). These techniques combine the efficiency of dynamic programming methods with the advantages of the state-of-the-art learning methods. Here we are interested in semi-supervised learning of structured label classification. An initial investigation of semi-supervised learning in the structured case is given in [2].

In discriminitive learning, one is interested in learning a mapping from an input $\mathbf{x} \in \mathcal{X}$ to an output or response $\mathbf{y} \in \mathcal{Y}$. In the multi-class case, this can be formulized by constructing a linear function of the form $F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$ and then mapping a given input x to the label $f(x)$ defined as follows.

$$f(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w}) \tag{1}$$

Here Ψ is a fixed feature map mapping the pair $(\mathbf{x}, \mathbf{y})$ into a linear vector space, $\mathbf{w}$ is a vector to be learned from training data. In the structured case, $\mathbf{x}$ and $\mathbf{y}$ are structured objects such as a sequence of words and a sequence of part of speech labels. In the semi-supervised case, we are to learn the vector w from a small labeled data set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_l, \mathbf{y}_l)\}$ as well as a large(r) unlabeled data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_u\}$. We adopt the common semi-supervised assumption that if two input patterns $\mathbf{x}$ and $\bar{\mathbf{x}}$ are similar then $f(\mathbf{x})$ and $f(\bar{\mathbf{x}})$ should also be similar. In the semi-supervised learning literature, this assumption is formalized by defining an adjacency graph over the unlabeled inputs and then defining a function on the nodes of that graph that is "smooth" in the sense that, to the largest extent possible, neighboring nodes are given similar labels. This approach must be modified when one needs to make an "out of sample" predication, i.e., to classify an input x which is not a node of the graph defined by the unlabeled training data. The out of sample problem is addressed by Belkin et al. by developing a linear discriminitive approach to semi-supervised learning. A linear discriminator in a predefined vector space can handle out of sample inputs. The unlabeled training data is handled by adding a regularization term which controls the smoothness of the labels over the adjacency graph on the unlabeled data. Here we generalize this approach to handle the structured case. The ability to overcome the out-of-sample problem is the main distinction between our work and [2].

Rather than define an adjaceny graph on structured inputs, we define an adjacency graph over the components of the structured inputs. We minimize an objective function that consists of three terms: the regularizer that controls the complexity of the discriminative function F, the regularizer that controls the smoothness of labels and the soft margin loss of labeled training examples where the minimum separation margin of $\mathbf{x}_i$ is $\gamma_i = F(\mathbf{x}_i, \mathbf{y}_i) - \max_{\mathbf{y} \neq \mathbf{y}_i} F(\mathbf{x}_i, \mathbf{y}) - \triangle(\mathbf{y}, \mathbf{y}_i)$. $\triangle$ is a linear function over cliques of $\mathbf{y}$ such as Hamming loss. Due to the linearity of F and the margin loss over cliques of $(\mathbf{x}, \mathbf{y})$ pairs, we obtain a quadratic program as in standard SVM, where the number of parameters scale polynomially with the size of $\mathbf{y}$. The optimal solution is stated in terms of the kernel values (which emerges from the complexity regularizer) of the components of both labeled and unlabeled data and the mixing parameters are functions of the graph kernel.

Defining the adjacency graph over the components of $\mathbf{x}$ effectively ignores (long range) dependencies of the components of the response variables corresponding to the unlabeled input patterns. This approach may be successful for applications where the components of $\mathbf{y}$ are mostly identifiable from the input patterns, thus where the correlation between the components of the response variables are not very strong. We investigate an alternative formulation for problems where correlations between the components of the response variables are stronger than the correlations between a component of the response variable and input patterns.

Topic: Learning algorithms
Preference: Poster

References

[1] Y. Altun. *Discriminative Methods for Label Sequence Learning.* PhD thesis, Department of Computer Science, Brown University, 2004.

[2] John Lafferty, Yan Liu, and Xiaojin Zhu. Kernel conditional random fields: Representation, clique selection, and semi-supervised learning. In *21th International Conference on Machine Learning (ICML)*, 2004.