Active Sample Selection and Correction Propagation on a Gradually-Augmented Graph

Hang Su[⋆†] Zhaozheng Yin[‡] Takeo Kanade[†] Seungil Huh[*]
[⋆] Department of Computer Science and Technology, Tsinghua Univelrsity
[†] Robotics Institute, Carnegie Mellon Univllersity
[‡] Department of Computer Science, Missouri University of Science and Technology
[*] Google

During the past decades, a lot of effort has been made on combining active and semi-supervised learning [3] [2], since they both try to maximize the utility of the expensive labeled data and exploit the abundant unlabeled data in real-world problem. However, most of the previous algorithms assume that queries are drawn from a closed pool and the characteristics of training and testing samples are the same, which may not be valid when the training and testing data are collected under different experimental conditions or characteristics of samples gradually change in a time-lapse sequence of data. Hence, it is worth to consider how to further incorporate human correction to achieve better label propagation results.

In order to find out which samples should be examined by human in order to maximize the return of investment or yield large accuracy improvement as early as possible, we derive a criterion that guides users to *actively* select error-prone samples by minimizing the expected prediction error of unlabeled data using the tool of transductive Rademacher complexity [1], which is defined as

Definition 1. *(**Transductive Rademacher Complexity**.) For a sample set $\mathcal{D} \triangleq \mathcal{L} \cup \mathcal{U} = \{\mathbf{x}_n\}_{n=1}^{N}$ with $N = N_l + N_u$, if $\mathcal{H}$ is a class of real-valued function on $\mathcal{D}$, the transductive Rademacher complexity of $\mathcal{H}$ is defined by generalizing the transductive Rademacher complexity to a multi-class version as*

$$\hat{R}(\mathcal{H};\mathcal{L}) = \left(\frac{1}{N_l} + \frac{1}{N_u}\right)\mathbb{E}_\sigma\left[\sup_{\mathbf{h}\in\mathcal{H}} \sum_{i=1}^{N_c} \sigma^T \mathbf{h}_i(\mathbf{X})\right], \tag{1}$$

where N_c is the number of classes, $\mathbf{h}_i(\mathbf{X}) = [h_i(\mathbf{x}_1);\cdots;h_i(\mathbf{x}_N)]$ is a column vector of hypothesis functions for the i_{th} class, and $\sigma = [\sigma_1,\cdots,\sigma_N]^T$ is a column vector of i.i.d. random variables such that σ_n is equal to 1 or -1 with the probability $p \in [0, \frac{1}{2}]$ for each, or 0 with the probability $1-2p$.

In this paper, we utilize the label propagation function $\mathbf{Y}_u^* = \mathbf{L}_{uu}^{-1}\mathbf{W}_{ul}\mathbf{Y}_l \triangleq \Gamma_{uu}\mathbf{W}_{ul}\mathbf{Y}_l$ as the hypothesis function, and p is set to $\frac{N_l N_u}{(Nl+Nu)^2}$. It has been proved that minimizing the bound of transductive Rademacher complexity is a proxy for minimizing an expected prediction error [1]. Afterwards, we propose a criterion by deriving the upper bound of the transductive Rademacher complexity for active sample selection as

Theorem 1. *(**Active sample selection criterion**.) The active sample selection for correction can be implemented as*

$$\mathcal{K}^* = \arg\min_{\mathcal{K}} \operatorname{tr}\left(\left(\mathbf{L}_{uu}\right)^{-2}\left(\mathbf{L}^2\right)_{uu}\right) \text{ with } u \in \mathcal{U}\setminus\mathcal{K}, \tag{2}$$

where $\mathbf{L}$ is the Laplacian matrix, $\mathcal{K}$ is a subset of $\mathcal{U}$ indicating samples selected for human examination.

Since the sample selection is an NP-hard problem, we propose a sequential minimization algorithm to find an optimal solution of $\mathbf{R}$; given that $k-1$ samples are already selected, the subsequent k_{th} sample is selected to result in the minimum increment of the objective function. Formally, the k_{th} sample is selected by solving the following problem:

$$k^* = \arg\min_k \left\{ \begin{array}{c} \operatorname{tr}(\Lambda^2\mathbf{R}_k^T\mathbf{R}_k) + 2\operatorname{tr}(\Phi_k\mathbf{R}_k^T\mathbf{R}_k\Lambda^2\mathbf{R}_k^T\mathbf{R}_k) \\ +\operatorname{tr}\left((\Phi_k\mathbf{R}_k^T\mathbf{R}_k)^2\Lambda^2\mathbf{R}_k^T\mathbf{R}_k\right) \end{array} \right\}, \tag{3}$$

where

$$\mathbf{R}_k^T\mathbf{R}_k = {\mathbf{R}_{k-1}}^T\mathbf{R}_{k-1} - \mathbf{q}_k\mathbf{q}_k^T,$$
$$\Phi_k = (\Phi_{k-1}^{-1} + \mathbf{q}_k\mathbf{q}_k^T)^{-1} = \Phi_{k-1} - \frac{\Phi_{k-1}\mathbf{q}_k\mathbf{q}_k^T\Phi_{k-1}}{1+\mathbf{q}_k^T\Phi_{k-1}\mathbf{q}_k}.$$

In the sequential minimization, given $\mathbf{R}_{k-1}$ and Φ_{k-1}, we are searching a column vector $\mathbf{q}_k$ (the transpose of the k_{th} row vector in $\mathbf{Q}$) that minimizes the objective function. Once the k_{th} sample is selected, $\mathbf{R}_k^T\mathbf{R}_k$ and Φ_k can be updated by substituting the optimal $\mathbf{q}_{k^*}$ into Eq. (4) and Eq. (4), respectively.

Once a user recognizes some errors when checking the samples recommended by the active sample selection, and corrects them manually, it is desirable to search for similar errors that can be fixed based on the given human intervention. Rather than rebuilding a statistical model from scratch using newly collected training data, we propose a scheme based on augmented graph [4] to handle a batch of samples at each round for more effective and efficient interaction.

Specifically, label propagation over this augmented graph can be obtained as

$$\mathbf{Y}_u^+ = \Gamma_{uu}^+\begin{bmatrix}\mathbf{W}_{ul} & \mathbf{W}_{us}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_l \\ \mathbf{Y}_s\end{bmatrix}, \tag{4}$$

where $\mathbf{W}_{us}$ is the submatrix indicating the relationship between the virtual supervisor and the unlabeled samples; $\mathbf{Y}_s$ is the label matrix of virtual supervisors; and Γ_{uu}^+ denotes inverse of the Laplacian submatrix of the augmented graph corresponding to unlabeled samples. The correction propagation is summarized as

Theorem 2. *(**Correction Propagation**.) The labels can be updated via the augmented graph as:*

$$\mathbf{Y}_u^+ = \mathbf{Y}_u + \Gamma_{uk}\Gamma_{kk}^{-1}(\mathbf{Y}_s - \mathbf{Y}_k), \tag{5}$$

where $\mathbf{Y}_u$ is the current label indicator, which has been updated during the previous correction propagation; $\mathbf{Y}_k$ is a submatix of $\mathbf{Y}_u$ that is constructed by stacking the rows of $\mathbf{Y}_u$ which correspond to samples verified by human. Γ_{kk} is a submatrix of the Laplacian matrix related to the human corrected samples, and Γ_{uk} is corresponding to the samples that are affected by the human corrections.

Hence, human corrections are propagated to the remaining unlabeled samples in $\mathcal{U}$ via $\Gamma_{uk}\Gamma_{kk}^{-1}$, therefore fixing samples undergoing similar errors.

Experimental results on both mismatched and time-evolved real-world data demonstrate that the human examination of the first 3% of samples results in approximately 10% accuracy improvement in the very early stage of human correction. This implies that the samples initially selected have typical errors, so correction on them can fix a lot of similar cases, thereby significantly reducing human efforts in refining the results.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[1] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

[2] Andrew B. Goldberg, Xiaojin Zhu, Alex Furger, and Jun-Ming Xu. OASIS: online active semi-supervised learning. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, Aug. 2011.

[3] Andrew Guillory and Jeff Bilmes. Active semi-supervised learning using submodular functions. In *Uncertainty in Artificial Intelligence (UAI)*, July 2011.

[4] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003.