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 University
‡ Department of Computer Science, Missouri University of Science and Technology
* Google

Abstract

When data have a complex manifold structure or the characteristics of data evolve over time, it is unrealistic to expect a graph-based semi-supervised learning method to achieve flawless classification given a small number of initial annotations. To address this issue with minimal human interventions, we propose (i) a sample selection criterion used for active query of informative samples by minimizing the expected prediction error, and (ii) an efficient correction propagation method that propagates human correction on selected samples over a gradually-augmented graph to unlabeled samples without rebuilding the affinity graph. Experimental results conducted on three real world datasets validate that our active sample selection and correction propagation algorithm quickly reaches high quality classification results with minimal human interventions.

1. Introduction

During the past decades, Graph-based Semi-Supervised Learning (GSSL) has many advances such as random walk [12] [9], manifold regularization [6], Gaussian Fields and Harmonic Functions (GFHF) [29], and Learning with Local and Global Consistency (LLGC) [28]. Comprehensive surveys of GSSL methods can be referred to [8] [30]. Since human labels are always expensive, it is expected to maximize the utility of labeled data and exploit the abundant unlabeled data in GSSL. Active learning provides a promising direction such that human interventions are guided via some criteria, e.g., uncertainty-based sampling [17] [21], density-based sampling [5] [7], expected error reduction [22] [16], etc. Settles gives a comprehensive survey in [20]. Ever since the seminal work of Zhu et al. [31], a lot of effort has been made on combining active and semi-supervised learning [14] [11]. There also exist several papers on active learning in conjunction with particularly graph-based SSL with different query strategies: maximizing graph cut size [15], learning optimal graph via active model selection [27], minimizing prediction label variance [18], and minimizing prediction error [13] [4].

It is noted that most of the aforementioned algorithms assume that queries are drawn from a closed pool and the characteristics of training and testing samples are the same. Unfortunately, this assumption may not be valid in many real-world scenarios because the training and testing data may be collected under different experimental conditions and therefore often exhibit differences in their statistics; and characteristics of samples may gradually change in a time-lapse sequence of data, e.g., online web-document classification. In this case, even with a well-constructed graph and a batch of informative samples as seeds, it is still unrealistic to expect that the label propagation results on a large pool of unlabeled data are error-free. Moreover, a classifier learned from initial labeling tends to result in more misclassifications over time until a proper correction mechanism is applied [24].

Hence, it is worth to consider how to further incorporate human correction to achieve better label propagation results. The intuitive way to conduct this would be to rebuild a statistical model from scratch using newly collected training data. Nevertheless, it is expensive or impossible to recollect the initial labeling set $\mathcal{L}$, because it requires a lot of human annotations or special devices/experiments to achieve the ground truth. It would be nice to reduce the efforts in re-collecting the data by re-using the previous labeled data and incorporating with subsequent human interventions, which leads to the following two questions:

1. How to find out which samples should be examined by human in order to maximize the return of investment,

*This research is partly supported by National Key Foundation R&D Projects No. 2013CB329403, National Natural Science Foundation of China (Nos. 61322308, 61332007, 61171172), NSF EPSCoR grant IIA-1355406 and NSF CAREER award IIS-1351049.

or yield large accuracy improvement as early as possible?

2. How to propagate human correction to unlabeled samples to automatically fix analogous errors for saving human efforts and speeding up the convergence to the best accuracy?

In this paper, we propose a sample selection and correction scheme to address these questions, and the main contributions are

- To derive a criterion guiding users to *actively* select error-prone samples by minimizing the expected prediction error of unlabeled data using the tool of transductive Rademacher complexity [10], and
- To propose a novel correction propagation matrix such that human correction on selected samples is *efficiently* propagated over an augmented graph rather than re-performing the complete label propagation from the beginning.

Sample selection and correction propagation are repeated until most of samples are classified into a specific class with high confidence, and the graph is *gradually* augmented along with human correction on informative samples.

The remainder of this paper is organized as follows. In Section 2, we elaborate the active sample selection algorithm for subsequential correction. In Section 3, we propose the correction propagation algorithm that fixes analogous errors. The experimental setup and results with discussions are presented in Section 4. Finally, the paper is summarized in Section 5.

2. Active Sample Selection for Correction

2.1. Overview of Graph-based Semi-Supervised Learning

Within the framework of Gaussian Fields and Harmonic Functions (GFHF) [29] which is proved in [26] to have comparable performance against the LLGC [28], we denote the annotated sample set as $\mathcal{L} \triangleq \{(\mathbf{x}_l, \mathbf{y}_l)\}_{l=1}^{N_l}$ where N_l is the number of annotated samples, and $\mathbf{x}_l$ and $\mathbf{y}_l$ are the feature vector and label vector of the l_{th} labeled sample, respectively. $\mathbf{y}_l$ is a row vector with only one 1 and 0s for the other entries such that $\mathbf{y}_l(k) = 1$ indicates that the sample belongs to the k_{th} class. In the same way, we define $\mathcal{U} \triangleq \{(\mathbf{x}_u, \mathbf{y}_u)\}_{u=1}^{N_u}$ with N_u being the number of unlabeled samples. The label vectors of unlabeled samples $\{\mathbf{y}_u\}$ are inferred via the label propagation procedure by minimizing the following objective function [19]:

$$f(\mathbf{Y}_u) = \text{tr}\Big([\mathbf{Y}_l; \mathbf{Y}_u]^T \mathbf{L} \begin{bmatrix} \mathbf{Y}_l \\ \mathbf{Y}_u \end{bmatrix}\Big), \qquad (1)$$

where $\mathbf{L}$ is the Laplacian matrix of the affinity graph; $\mathbf{Y}_l$ and $\mathbf{Y}_u$ are the label matrices of labeled and unlabeled samples constructed by stacking up the indicator vectors $\{\mathbf{y}_l\}_{l=1}^{N_l}$ and $\{\mathbf{y}_u\}_{u=1}^{N_u}$, respectively. Differentiating $f(\mathbf{Y}_u)$ in Eq. (1) with respect to $\mathbf{Y}_u$ and setting it to zero yield the closed-form solution on the indicator for unlabeled samples as

$$\mathcal{H} : \mathbf{Y}_u^* = \mathbf{L}_{uu}^{-1} \mathbf{W}_{ul} \mathbf{Y}_l \triangleq \mathbf{\Gamma}_{uu} \mathbf{W}_{ul} \mathbf{Y}_l, \qquad (2)$$

where $\mathbf{W}_{ul}$ characterizes the pairwise similarity between unlabeled and labeled samples with $\mathbf{W}_{ul} = -\mathbf{L}_{ul}$; $\mathbf{L}_{uu}$ denotes the submatrix corresponding to unlabeled samples in $\mathbf{L}$. We denote the inverse of $\mathbf{L}_{uu}$ as $\mathbf{\Gamma}_{uu}$, i.e., $\mathbf{\Gamma}_{uu} = \mathbf{L}_{uu}^{-1}$. Note that the elements of $\mathbf{Y}_u^*$ are not binary but real numbers, thus $\mathbf{Y}_u^*$ can be considered as soft label results. The hard label vector can be obtained simply by converting the maximum value in each $\mathbf{y}_u^*$ into 1 and the others into 0.

2.2. Active Sample Selection by Minimizing the Expected Prediction Error

It is expected that there exist misclassifications after label propagation, but it is not easy for a user to determine which samples to check and correct if needed in order to achieve higher accuracy. In this section, we propose a strategy to select samples that are likely to minimize the expected prediction errors using the tool of transductive Rademacher complexity [10].

2.2.1 Bound of Transductive Rademacher Complexity

We generalize the transductive Rademacher complexity to a multi-class version 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 as*

$$\hat{R}(\mathcal{H}; \mathcal{L}) = \Big(\frac{1}{N_l} + \frac{1}{N_u}\Big)\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{h} \in \mathcal{H}} \sum_{i=1}^{N_c} \boldsymbol{\sigma}^T \mathbf{h}_i(\mathbf{X})\Big], \qquad (3)$$

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 $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_N]^T$ 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 in Eq. (2) 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 [10]. In the following, we derive the upper bound of the transductive Rademacher complexity in Theorem 1, which serves as the theoretical foundation of our proposed algorithm.

Theorem 1. ***(Bound of transductive Rademacher complexity**.) The transductive Rademacher complexity of label propagation based on $\mathcal{H}$ in Eq.* (2) *is upper bounded by*

$$\hat{\mathcal{R}}(\mathcal{H};\mathcal{L}) \leq C\sqrt{\text{tr}(\mathbf{P}\mathbf{P}^T)}, \tag{4}$$

where $\mathbf{P}$ is the label propagation matrix $\mathbf{P} = \mathbf{\Gamma}_{uu}\mathbf{W}_{ul}$, tr *is the trace operator, and C is a constant.*

Proof. In Eq. (2), the label propagation is given by

$$\mathbf{Y}_u = \mathbf{\Gamma}_{uu}\mathbf{W}_{ul}\mathbf{Y}_l \triangleq \mathbf{P}\mathbf{Y}_l. \tag{5}$$

By definition 1, the transductive Rademacher complexity for label propagation is computed as

$$\hat{\mathcal{R}}(\mathcal{H};\mathcal{L}) = (\frac{1}{N_l}+\frac{1}{N_u})\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{h}\in\mathcal{H}}\sum_{i=1}^{N_c}\boldsymbol{\sigma}^T\mathbf{P}\mathbf{Y}_l^{(i)}\Big], \tag{6}$$

where $\mathbf{Y}_l^{(i)}$ is the i_{th} column of $\mathbf{Y}_l$, i.e., $\mathbf{Y}_l^{(i)}(n) = 1$ indicates that the n_{th} sample belongs to the i_{th} class, otherwise $\mathbf{Y}_l^{(i)}(n) = 0$. Therefore, $\mathbf{P}\mathbf{Y}_l^{(i)}$ indicates the probability of samples belonging to the i_{th} class. By applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{H};\mathcal{L}) &= (\frac{1}{N_l}+\frac{1}{N_u})\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{h}\in\mathcal{H}}\boldsymbol{\sigma}^T\mathbf{P}\sum_{i=1}^{N_c}\mathbf{Y}_l^{(i)}\Big]\\ &\leq (\frac{1}{N_l}+\frac{1}{N_u})\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{h}\in\mathcal{H}}\|\boldsymbol{\sigma}^T\mathbf{P}\|_2\|\sum_{i=1}^{N_c}\mathbf{Y}_l^{(i)}\|_2\Big]\\ &= (\frac{1}{N_l}+\frac{1}{N_u})\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{h}\in\mathcal{H}}\|\boldsymbol{\sigma}^T\mathbf{P}\|_2\|\mathbf{Y}_l\|_2\Big]. \end{aligned}\tag{7}$$

Since there are N_l labeled samples, $\|\mathbf{Y}_l\|_2 = \sqrt{N_l}$. Using the property of inner product, we have

$$\begin{aligned}&\hat{\mathcal{R}}(\mathcal{H};\mathcal{L}) \leq (\tfrac{1}{N_l}+\tfrac{1}{N_u})\sqrt{N_l}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sqrt{\boldsymbol{\sigma}^T\mathbf{P}\mathbf{P}^T\boldsymbol{\sigma}}\Big] \qquad (8)\\ &= (\tfrac{1}{N_l}+\tfrac{1}{N_u})\sqrt{N_l}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sqrt{\sum_{i,j=1}^{N_l+N_u}\sigma_i\sigma_j\langle\mathbf{P}(i,:),\mathbf{P}(j,:)\rangle}\Big].\end{aligned}$$

Using the Jensen inequalities, the expectation term in Eq. (8) is upper bounded by

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sqrt{\sum_{i,j=1}^{N_l+N_u}\sigma_i\sigma_j\langle\mathbf{P}(i,:),\mathbf{P}(j,:)\rangle}\Big]\\ \leq\ &\sqrt{\sum_{i,j=1}^{N_l+N_u}\mathbb{E}_{\boldsymbol{\sigma}}\left\{\sigma_i\sigma_j\langle\mathbf{P}(i,:),\mathbf{P}(j,:)\rangle\right\}}\\ =\ &\sqrt{\sum_{i=1}^{N_l+N_u}\tfrac{2N_lN_u}{(N_l+N_u)^2}\langle\mathbf{P}(i,:),\mathbf{P}(i,:)\rangle}\end{aligned}\tag{9}$$

Note that Eq. (9) is obtained by the expectation of Rademacher variables $\boldsymbol{\sigma}$ in definition 1. By substituting Eq. (9) into Eq (8), we can obtain the transductive Rademacher complexity is upper bounded by

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{H};\mathcal{L}) &\leq \sqrt{\frac{2}{N_u}}\left[\sqrt{\sum_{i=1}^{N_l+N_u}\langle\mathbf{P}(i,:),\mathbf{P}(i,:)\rangle}\right] \qquad (10)\\ &= C\sqrt{\text{tr}(\mathbf{P}\mathbf{P}^T)}. \qquad (11)\end{aligned}$$

Hereby, we obtain the upper bound of the transductive Rademacher complexity in Eq. (4) with $C = \sqrt{\frac{2}{N_u}}$. □

2.2.2 Sample Selection Criterion

Based on the bound in Theorem 1, we derive a criterion for active sample selection:

Theorem 2. ***(Active sample selection criterion**.) The active sample selection for correction can be implemented as*

$$\mathcal{K}^* = \arg\min_{\mathcal{K}} \text{tr}\left(\left(\mathbf{L}_{uu}\right)^{-2}\left(\mathbf{L}^2\right)_{uu}\right) \text{ with } u \in \mathcal{U}\setminus\mathcal{K}, \tag{12}$$

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

Proof. Minimizing the upper bound of the transductive Rademacher complexity in Theorem 1 is equivalent to minimizing $\text{tr}(\mathbf{P}\mathbf{P}^T)$ as we can ignore the constant factor and square root function in Eq. (4). Applying the cyclic property of trace gives

$$\begin{aligned}\text{tr}(\mathbf{P}\mathbf{P}^T) &= \text{tr}(\mathbf{L}_{uu}^{-1}\mathbf{W}_{ul}\mathbf{W}_{ul}^T\mathbf{L}_{uu}^{-1})\\ &= \text{tr}\left((\mathbf{L}_{uu})^{-1}(\mathbf{L}_{uu})^{-1}\mathbf{W}_{ul}\mathbf{W}_{lu}\right). \qquad (13)\end{aligned}$$

Note that $\mathbf{W}_{ul}\mathbf{W}_{lu} = (\mathbf{L}^2)_{uu} - \mathbf{L}_{uu}\mathbf{L}_{uu}$ because

$$\begin{aligned}(\mathbf{L}^2)_{uu} &= \left(\begin{bmatrix}\mathbf{L}_{ll} & \mathbf{L}_{lu}\\ \mathbf{L}_{ul} & \mathbf{L}_{uu}\end{bmatrix}\begin{bmatrix}\mathbf{L}_{ll} & \mathbf{L}_{lu}\\ \mathbf{L}_{ul} & \mathbf{L}_{uu}\end{bmatrix}\right)_{uu}\\ &= \mathbf{L}_{ul}\mathbf{L}_{lu} + \mathbf{L}_{uu}\mathbf{L}_{uu}\\ &= \mathbf{W}_{ul}\mathbf{W}_{lu} + \mathbf{L}_{uu}\mathbf{L}_{uu}. \qquad (14)\end{aligned}$$

By putting this into Eq. (13), we obtain

$$\begin{aligned}\text{tr}(\mathbf{P}\mathbf{P}^T) &= \text{tr}\Big((\mathbf{L}_{uu})^{-1}(\mathbf{L}_{uu})^{-1}\big((\mathbf{L}^2)_{uu} - \mathbf{L}_{uu}\mathbf{L}_{uu}\big)\Big)\\ &= \text{tr}\left((\mathbf{L}_{uu})^{-2}(\mathbf{L}^2)_{uu}\right) - N_u. \qquad (15)\end{aligned}$$

Thus minimizing $\text{tr}(\mathbf{P}\mathbf{P}^T)$ is equivalent to minimizing $\text{tr}\left((\mathbf{L}_{uu})^{-2}(\mathbf{L}^2)_{uu}\right)$. □

Solving Eq. (12) selects samples into $\mathcal{K}^*$ such that the upper bound of the transductive Rademacher complexity on the rest, i.e., $\mathcal{U}\setminus\mathcal{K}^*$, is the minimum, meaning that $\mathcal{K}^*$ includes the most informative samples out of $\mathcal{U}$ for human examination.

2.2.3 Sequential Optimization to Select Informative Samples

Since the active sample selection criterion in Theorem 2 is a combinatorial optimization problem, finding the global optimal solution is an NP-hard problem. To solve it, we form a binary matrix $\bar{\mathbf{S}} = [\bar{s}_{ij}]$ sized $N \times N_u$ such that $\bar{s}_{ij} = 1$ if the i_{th} sample in $\mathcal{D}$ remains unlabeled after human examination and corresponds to the j_{th} samples in $\mathcal{U}$, and 0 otherwise. Then $\mathbf{L}_{uu} = \bar{\mathbf{S}}^T \mathbf{L} \bar{\mathbf{S}}$, so the problem in Theorem 2 can be reformulated as

$$\bar{\mathbf{S}}^* = \arg\min_{\bar{\mathbf{S}}} \text{tr}\left(\left(\bar{\mathbf{S}}^T \mathbf{L} \bar{\mathbf{S}} \right)^{-2} \bar{\mathbf{S}}^T \mathbf{L}^2 \bar{\mathbf{S}} \right), \quad (16)$$

$$\text{s.t.} \begin{cases} \bar{\mathbf{S}} \in \{0,1\}^{N \times N_u}, \\ \bar{\mathbf{S}}^T \bar{\mathbf{S}} = \mathbf{I}_{N_u} \end{cases} \quad (17)$$

Using eigenvalue analysis, the Laplacian matrix is decomposed as $\mathbf{L} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$ where $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues and $\mathbf{Q}$ is a matrix containing normalized eigenvectors. Due to the orthogonality of eigenvectors (i.e., $\mathbf{Q}\mathbf{Q}^T = I_N$), the problem in Eq. (16) can be rewritten as

$$\bar{\mathbf{S}}^* = \arg\min_{\bar{\mathbf{S}}} \text{tr}\left((\bar{\mathbf{S}}^T \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \bar{\mathbf{S}})^{-2} (\bar{\mathbf{S}}^T \mathbf{Q} \mathbf{\Lambda}^2 \mathbf{Q}^T \bar{\mathbf{S}}) \right). \quad (18)$$

Denoting $\mathbf{R} = \bar{\mathbf{S}}^T \mathbf{Q}$, the problem can be rewritten as

$$\mathbf{R}^* = \arg\min_{\mathbf{R}} \text{tr}\left(\left(\mathbf{R} \mathbf{\Lambda} \mathbf{R}^T \right)^{-2} \left(\mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T \right) \right). \quad (19)$$

Since $\mathbf{R}\mathbf{R}^T = \mathbf{I}_{N_u}$, we have

$$\begin{aligned} \left(\mathbf{R} \mathbf{\Lambda} \mathbf{R}^T \right)^{-1} &= \left(-\mathbf{I} + \mathbf{R}(\mathbf{\Lambda} + \mathbf{I}) \mathbf{R}^T \right)^{-1} \\ &= -\mathbf{I} - \mathbf{R} \left((\mathbf{\Lambda} + \mathbf{I})^{-1} - \mathbf{R}^T \mathbf{R} \right)^{-1} \mathbf{R}^T \\ &\triangleq -\mathbf{I} - \mathbf{R} \mathbf{\Phi} \mathbf{R}^T \end{aligned} \quad (20)$$

where $\mathbf{\Phi} = \left((\mathbf{\Lambda} + \mathbf{I})^{-1} - \mathbf{R}^T \mathbf{R} \right)^{-1}$. Therefore, we can obtain

$$\begin{aligned} & \text{tr}\left(\left(\mathbf{R} \mathbf{\Lambda} \mathbf{R}^T \right)^{-2} \left(\mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T \right) \right) \quad (21) \\ = \; & \text{tr}\left(\left(-\mathbf{I} - \mathbf{R} \mathbf{\Phi} \mathbf{R}^T \right)^2 \left(\mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T \right) \right) \\ = \; & \text{tr}\left(\left(\mathbf{I} + 2\mathbf{R} \mathbf{\Phi} \mathbf{R}^T + \mathbf{R} \mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Phi} \mathbf{R}^T \right) \left(\mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T \right) \right) \end{aligned}$$

It is easy to see that Eq. (21) can be expressed as a sum of three trace terms, which are

$$\text{tr}(\mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T)) = \text{tr}(\mathbf{\Lambda}^2 \mathbf{R}^T \mathbf{R}), \quad (22)$$

$$2\text{tr}(\mathbf{R} \mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T) = 2\text{tr}(\mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T \mathbf{R}), \quad (23)$$

and

$$\text{tr}(\mathbf{R} \mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T) = \text{tr}(\mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Phi} \mathbf{R}^T \mathbf{R} \mathbf{\Lambda}^2 \mathbf{R}^T \mathbf{R}). \quad (24)$$

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:

$$k^* = \arg\min_k \left\{ \begin{array}{c} \text{tr}(\mathbf{\Lambda}^2 \mathbf{R}_k^T \mathbf{R}_k) + 2\text{tr}(\mathbf{\Phi}_k \mathbf{R}_k^T \mathbf{R}_k \mathbf{\Lambda}^2 \mathbf{R}_k^T \mathbf{R}_k) \\ +\text{tr}\left((\mathbf{\Phi}_k \mathbf{R}_k^T \mathbf{R}_k)^2 \mathbf{\Lambda}^2 \mathbf{R}_k^T \mathbf{R}_k \right) \end{array} \right\}, \quad (25)$$

where

$$\mathbf{R}_k^T \mathbf{R}_k = {\mathbf{R}_{k-1}}^T \mathbf{R}_{k-1} - \mathbf{q}_k \mathbf{q}_k^T, \quad (26)$$

$$\mathbf{\Phi}_k = (\mathbf{\Phi}_{k-1}^{-1} + \mathbf{q}_k \mathbf{q}_k^T)^{-1} = \mathbf{\Phi}_{k-1} - \frac{\mathbf{\Phi}_{k-1} \mathbf{q}_k \mathbf{q}_k^T \mathbf{\Phi}_{k-1}}{1 + \mathbf{q}_k^T \mathbf{\Phi}_{k-1} \mathbf{q}_k} \quad (27)$$

Note that $\mathbf{\Phi}_k$ can be updated by matrix (vector) multiplication and addition in Eq. (27), which is much more efficient than the matrix inverse.

In the sequential minimization, given $\mathbf{R}_{k-1}$ and $\mathbf{\Phi}_{k-1}$, 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 $\mathbf{\Phi}_k$ can be updated by substituting the optimal $\mathbf{q}_{k^*}$ into Eq. (26) and Eq. (27), respectively. We summarize the sample selection method in Algorithm 1, following which a batch of informative samples can be selected for human correction.

Algorithm 1 Active Sample Selection by Minimizing Transductive Rademacher Complexity

1: **Input**: Number of samples to select N_s, $\mathbf{R}_1$ and $\mathbf{\Phi}_1$.
2: **Output**: Actively selected sample set $\mathcal{K}$
3: **for** $k = 1 \rightarrow N_s$ **do**
4: Select a sample out of unlabeled samples by solving

$$k^* = \arg\min_k \left\{ \begin{array}{c} \text{tr}(\mathbf{\Lambda}^2 \mathbf{R}_k^T \mathbf{R}_k) + 2\text{tr}(\mathbf{\Phi}_k \mathbf{R}_k^T \mathbf{R}_k \mathbf{\Lambda}^2 \mathbf{R}_k^T \mathbf{R}_k) \\ +\text{tr}\left((\mathbf{\Phi}_k \mathbf{R}_k^T \mathbf{R}_k)^2 \mathbf{\Lambda}^2 \mathbf{R}_k^T \mathbf{R}_k \right) \end{array} \right\}.$$

5: Update the selected sample set: $\mathcal{K} \leftarrow \mathcal{K} \cup \{k^*\}$.
6: $\mathbf{R}_k^T \mathbf{R}_k \leftarrow {\mathbf{R}_{k-1}}^T \mathbf{R}_{k-1} - \mathbf{q}_{k^*} \mathbf{q}_{k^*}^T$; and

$$\mathbf{\Phi}_k \leftarrow \mathbf{\Phi}_{k-1} - \frac{\mathbf{\Phi}_{k-1} \mathbf{q}_{k^*} \mathbf{q}_{k^*}^T \mathbf{\Phi}_{k-1}}{1 + \mathbf{q}_{k^*}^T \mathbf{\Phi}_{k-1} \mathbf{q}_{k^*}}$$

7: **end for**

Note: $\mathbf{R}_1$ is initialized by stacking the rows in $\mathbf{Q}$ that are corresponding to samples in $\mathcal{U}$ before human correction is conducted, $\mathbf{\Phi}_1 = \left((\mathbf{\Lambda} + \mathbf{I})^{-1} - \mathbf{R}_1^T \mathbf{R}_1 \right)^{-1}$, and $\mathcal{K}$ is $\emptyset$ as initialization.

3. Correction Propagation on a Gradually-Augmented Graph

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. The intuitive way to conduct this task would be to rebuild or modify the affinity matrix in Eq. (1) and re-perform label propagation as Eq. (2). However, this re-propagation scheme is very inefficient for the interactive correction because the inverse of the large Laplacian matrix in Eq. (2) needs to be recalculated again and again. Instead, we propose a scheme based on augmented graph [31] to handle a batch of samples at each round for more effective and efficient interaction.

We build an *augmented graph* by adding auxiliary nodes (called *virtual supervisors* and denoted by $\mathcal{S} = \{\mathbf{y}_s\}_{s=1}^{N_s}$ with N_s being the number of virtual supervisors and $\mathbf{y}_s$ being the given human label on the s-th sample) and connecting them to the corrected samples with weight w such that $w \to +\infty$. Then the weighted adjacency matrix of the augmented graph can be built by adding some rows and columns to the original weighted adjacency matrix as follows:

$$\mathbf{W}^+ \leftarrow \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} & \mathbf{O}_{ls} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} & \mathbf{W}_{us} \\ \mathbf{O}_{sl} & \mathbf{W}_{su} & \mathbf{O}_{ss} \end{bmatrix}, \tag{28}$$

where $\mathbf{O}_{ls}$, $\mathbf{O}_{sl}$ and $\mathbf{O}_{ss}$ are zero submatrices with appropriate sizes, and $\mathbf{W}_{us} = w\mathbf{Z}_{us}$ and $\mathbf{W}_{su} = w\mathbf{Z}_{su} = w\mathbf{Z}_{us}^T$ where $\mathbf{Z}_{us}$ and $\mathbf{Z}_{su}$ are binary indicator matrices indicating which virtual supervisor is connected to which unlabeled sample.

Similar as Eq. (1), label propagation over this augmented graph can be obtained as

$$\begin{aligned} \mathbf{Y}_u^+ &= (\mathbf{L}_{uu}^+)^{-1} \begin{bmatrix} \mathbf{W}_{ul} & \mathbf{W}_{us} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_l \\ \mathbf{Y}_s \end{bmatrix} \\ &= \mathbf{\Gamma}_{uu}^+ (\mathbf{W}_{ul}\mathbf{Y}_l + w\mathbf{Z}_{us}\mathbf{Y}_s) \end{aligned} \tag{29}$$

where $\mathbf{L}_{uu}^+$ denotes the Laplacian submatrix of the augmented graph corresponding to unlabeled samples and $\mathbf{\Gamma}_{uu}^+$ is its inverse; $\mathbf{Y}_s$ is the label matrix of virtual supervisors.

In the following, we derive how to efficiently compute $\mathbf{\Gamma}_{uu}^+$ and $\mathbf{Y}_u^+$. The submatrix of the degree matrix of the augmented graph corresponding to unlabeled samples is computed as

$$\begin{aligned} \mathbf{D}_{uu}^+(i,i) &= \sum_{j=1}^{N_l} \mathbf{W}_{ul}(i,j) + \sum_{j=1}^{N_u} \mathbf{W}_{uu}(i,j) + \sum_{j=1}^{N_s} \mathbf{W}_{us}(i,j), \\ &= \mathbf{D}_{uu}(i,i) + w(\mathbf{Z}_{us}\mathbf{Z}_{su})(i,i). \end{aligned} \tag{30}$$

Note that $\sum_{j=1}^{N_s} \mathbf{Z}_{us}(i,j) = (\mathbf{Z}_{us}\mathbf{Z}_{su})(i,i)$ since $\mathbf{Z}_{us}$ is a binary indicator matrix. Then, $\mathbf{L}_{uu}^+$ can be computed as

$$\mathbf{L}_{uu}^+ = \mathbf{D}_{uu}^+ - \mathbf{W}_{uu}^+ = \mathbf{L}_{uu} + w\mathbf{Z}_{us}\mathbf{Z}_{su}, \tag{31}$$

and accordingly $\mathbf{\Gamma}_{uu}^+$ can be calculated using the Binomial inverse theorem [25] as

$$\begin{aligned} \mathbf{\Gamma}_{uu}^+ &= (\mathbf{L}_{uu}^+)^{-1} = (\mathbf{L}_{uu} + w\mathbf{Z}_{us}\mathbf{Z}_{su})^{-1} \\ &= \mathbf{L}_{uu}^{-1} - w\mathbf{L}_{uu}^{-1}\mathbf{Z}_{us}(\mathbf{I}_{N_s} + w\mathbf{Z}_{su}\mathbf{L}_{uu}^{-1}\mathbf{Z}_{us})^{-1}\mathbf{Z}_{su}\mathbf{L}_{uu}^{-1} \\ &= \mathbf{\Gamma}_{uu} - \mathbf{\Gamma}_{uk}(\mathbf{I}_{N_s}/w + \mathbf{\Gamma}_{kk})^{-1}\mathbf{\Gamma}_{ku} \end{aligned} \tag{32}$$

where $\mathbf{\Gamma}_{uk}$, $\mathbf{\Gamma}_{kk}$, and $\mathbf{\Gamma}_{ku}$ are submatrices of $\mathbf{\Gamma}_{uu}$ such that $\mathbf{\Gamma}_{uk} = \mathbf{\Gamma}_{uu}\mathbf{Z}_{us} = [\mathbf{\Gamma}_{uu}(i,j)], \forall i \in \mathcal{U}, j \in \mathcal{K}$; $\mathbf{\Gamma}_{kk} = [\mathbf{\Gamma}_{uu}(i,j)], \forall i \in \mathcal{K}, j \in \mathcal{K}$; and $\mathbf{\Gamma}_{ku} = \mathbf{Z}_{su}\mathbf{\Gamma}_{uu} = [\mathbf{\Gamma}_{uu}(i,j)], \forall i \in \mathcal{K}, j \in \mathcal{U}$ where $\mathcal{K}$ is a subset of selected samples among $\mathcal{U}$ for human verification (Section 2).

By substituting Eq. (32) into Eq. (29), the updated labels can be obtained by some linear algebra computations:

$$\mathbf{Y}_u^+ = \mathbf{Y}_u + \mathbf{\Gamma}_{uk}\mathbf{\Gamma}_{kk}^{-1}(\mathbf{Y}_s - \mathbf{Y}_k), \tag{33}$$

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. $\mathbf{\Gamma}_{kk}$ is related to the human corrected samples, and $\mathbf{\Gamma}_{uk}$ 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 $\mathbf{\Gamma}_{uk}\mathbf{\Gamma}_{kk}^{-1}$, therefore fixing samples undergoing similar errors.

We denote $\mathbf{\Gamma}_{uk}\mathbf{\Gamma}_{kk}^{-1}$ as the *correction propagation matrix*. By propagating the human correction information ($\mathbf{Y}_s$) through this matrix, the classification results can be incrementally improved. Note that, to update $\mathbf{Y}_u^+$, we compute the inverse of $\mathbf{\Gamma}_{kk}$ in Eq. (33) which is a small $N_s \times N_s$ matrix, rather than $\mathbf{L}_{uu}^+$ in Eq. (29) which is a large $N_u \times N_u$ matrix. As a result, correction propagation is efficiently performed, and misclassification is effectively corrected in the meantime. In summary, we detail the active correction propagation algorithm in Algorithm 2.

Algorithm 2 Active Correction Propagation

1: **Input**: Number of corrected samples N_s.
2: **Output**: Label matrix of unlabeled samples $\mathbf{Y}_u$
3: **repeat**
4: Use Algorithm 1 to select N_s samples as subset $\mathcal{K}$, and obtain its human correction $\mathbf{Y}_s$.
5: Implement correction propagation as
$$\mathbf{Y}_u \leftarrow \mathbf{Y}_u + \mathbf{\Gamma}_{uk}\mathbf{\Gamma}_{kk}^{-1}(\mathbf{Y}_s - \mathbf{Y}_k).$$
6: Compute the sample uncertainty using entropy as
$$H(u) = -\sum_{i=1}^{N_c} \mathbf{Y}_u^{(i)^T} \log \mathbf{Y}_u^{(i)}.$$
7: **until** most labels are certain, i.e., $\sum_{u=1}^{N_u} H(u) < th$.

4. Experiments

Our method is evaluated and compared with several other learning methods extensively on several datasets including mismatch between training and testing data, and data evolution over time.

4.1. Comparison Methods

To verify the effectiveness of our active sample selection and correction propagation in classification, we compare our method against alternative learning algorithms. First, we implemented two classification algorithms solely based on initially annotated samples without subsequent human correction:

- **Random annotation and no correction [29]**: Random annotation performs uniform sampling on the dataset to select samples for initial labeling. The classification is achieved by Eq. (2) without the correction step, which acts as the baseline method.
- **Active annotation and no correction [13]**: We apply the same label propagation algorithm by using the most informative samples as seeds that are drawn actively by minimizing the expected prediction error, but without any correction.

We also implemented three algorithms that apply correction mechanism on top of active annotation at the beginning:

- **Active annotation and random correction**: We select a portion of the most informative samples via active annotation [13] as initial labeling, then randomly select samples and request human checks and corrections. Manual interventions are propagated to the unlabeled samples following Eq. (33).
- **Active annotation and sequential correction**: As a comparison, we implement the sequential correction [31] on top of the samples that are actively selected via active annotation [13].
- **Active annotation and active correction**: After obtaining a portion of samples via active annotation [13], we actively correct the results based on our proposed sample selection and correction propagation algorithms.

As a special case, we implement the classification task by ignoring the labeled data; then, our active sample selection and subsequential correction propagation becomes

- **Active correction and no annotation**: In section 2.2.3, when $\mathcal{D} = \mathcal{U}$ (i.e., no initial annotations), $\bar{\mathbf{S}} = \mathbf{I}_N$ and $\mathbf{R}_{k-1}$ is initialized as $\mathbf{Q}$ in Algorithm 1. In this scenario, we actively select the informative samples and then utilize them to conduct the classification.

In order to reduce the bias, the result is averaged over 10 trials on the entire dataset if samples are selected randomly, including *random annotation and no correction* and *active annotation and random correction*.

4.2. Classification on Mismatched Data

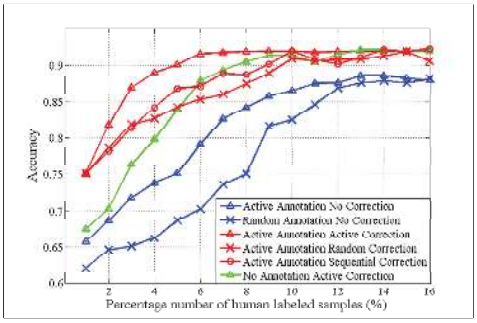
In order to show the performance of our proposed algorithm on data whose training and testing samples are of different feature distributions, i.e., feature related but not matched very well, we use two real-world benchmark datasets:

- **USPS handwritten digits (USPS)**. This database has a set of images containing handwritten digits, which are divided into training and testing sets with 4649 samples each [1].
- **ISOLET spoken letter (ISOLET).** This database contains 150 subjects who spoke the name of each letter of the alphabet twice [2], and we use the first 60 subjects as training data and others as testing data.

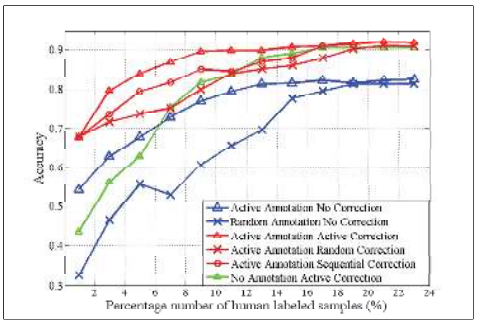
To further enforce the mismatch between the training and testing data, we artificially introduce errors to the labels and add noise to the feature in the training dataset. Specifically, 5% of labels in the training dataset are chosen randomly, then intentionally changed to wrong labels; and we added Gaussian noise to the feature vector $\mathbf{x}$ with zero mean and standard deviation $\sigma = 10\% \cdot \max(\|\mathbf{x}\|_2)$.

The classification results on the testing dataset are reported in Fig. 1. As is expected and observed, the accuracies (y-axis) of all methods including the baseline method are improved as the number of labeled samples (x-axis) increases. If the samples are actively selected for labeling from the training dataset, the performance improves more rapidly and levels off earlier than random annotation (blue triangle vs. blue X mark), since the informative samples are drawn early and annotated first. However, the performance is not further improved at some low point once the number of the annotated samples reaches a certain level due to the mismatching between the testing and training data (blue curves).

In contrast, we can achieve a higher performance when correction is involved. We reuse a portion of the most informative samples that are actively drawn from the training dataset, and apply correction propagation on top of the initial label propagation results. If the correction is conducted randomly (red curves marked with X marks), the classification accuracy improves more rapidly than active annotation without correction (blue curves marked with triangle marks), because correction is propagated to other samples, automatically correcting similar errors. However, in this *active annotation and random correction*, users have no clue on which samples are informative to be checked and corrected if needed.



(a) USPS



(b) ISOLET

Figure 1: Classification accuracy vs. the number of human labels. The blue, red and green curves represent the classification are implemented by annotation, annotation and correction, and solely correction, respectively.

Our *active annotation and active correction* provides guidance for human to examine the most informative samples. Its accuracy (red curves marked with triangle marks) demonstrates that the performance converges to a high quality result more rapidly than the alternative methods thanks to active selection of informative samples that minimizes the expected predicted error of unlabeled samples. The human examination of the first 3% of samples results in approximately 10% accuracy improvement, as shown in the very early stage of human correction (beginning of red curves marked with triangle marks). 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. On contrast, the sequential correction tries to draw samples by minimizing the risk of harmonic energy function, but it does not guarantee the quality of predictions on the unlabeled data, which results in a slower convergence rate (red curves with circle marks).

As a special case, we ignored the samples that are drawn from the training data, and the classification is implemented solely based on correction (green curve with triangle marks). As is observed, the performance is not comparable to the previous algorithms that involve active annotation when the number of the corrected sample is small, since it does not leverage the information from labeled examples. The curve is getting closer to that with *active annotation and active correction*, since the impact of corrected examples from the testing dataset tends to dominate. However, the convergence speed is still slower than that of *active annotation and active correction*.

4.3. Classification on Gradually Evolved Data

Our proposed method also shows advantages on a dataset that expands over time with more and more unseen data. In order to demonstrate the empirical evidence, we use the time-lapse image sequences for evaluation.

- **Time-lapse phase contrast microscopy images (Cell).** This database contains different types of cells, including muscle stem cell of a progeroid mouse (Seq1) and C2C12 myoblastic stem cell (Seq2). Each frame contains as few as 50 cells and as many as 800 cells. From each sequence, we select image frames with a certain interval [3]. Phase retardation feature of cells is restored, and each frame is partitioned into phase-homogeneous atoms [23]. Cell segmentation is realized by classifying the atoms into specific classes.

Specifically, we annotated the cells in the first frame of each sequence, based on which we train a classifier that is applied for segmentation of cells in the subsequent frames. As various factors from overall illumination to each cell's visual properties adapted to the environment (e.g., increased density of cells) can change over time, a gradually increasing level of misclassification is expected. In such a case, misclassifications are fixed based on our proposed correction propagation method in Algorithm 2.

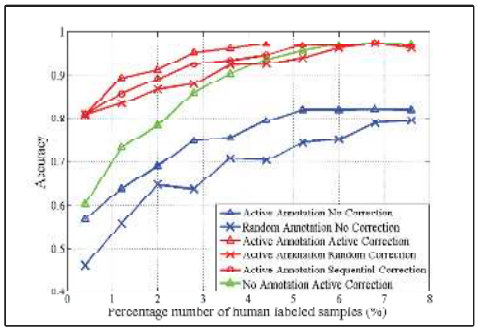


Figure 3: Classification accuracy vs. the number of human labels on Cell.

Examples of the cell segmentation results for the subsequential frames are shown in Fig. 2. In column (c) and (d), we demonstrate the soft and hard segmentation results based on label propagation of the initial human annotation from the first image. As can be seen in column (d), there contain several errors in the segmentation result, i.e., some dark cells are missed and bright halos are misclassified into cells in Seq1; and some dark-adjacent atoms are classified into dark cells in Seq2. These errors are partly ascribed to inadequate or unbalance human labels, but also to visual changes of cells and environment over time. Some of

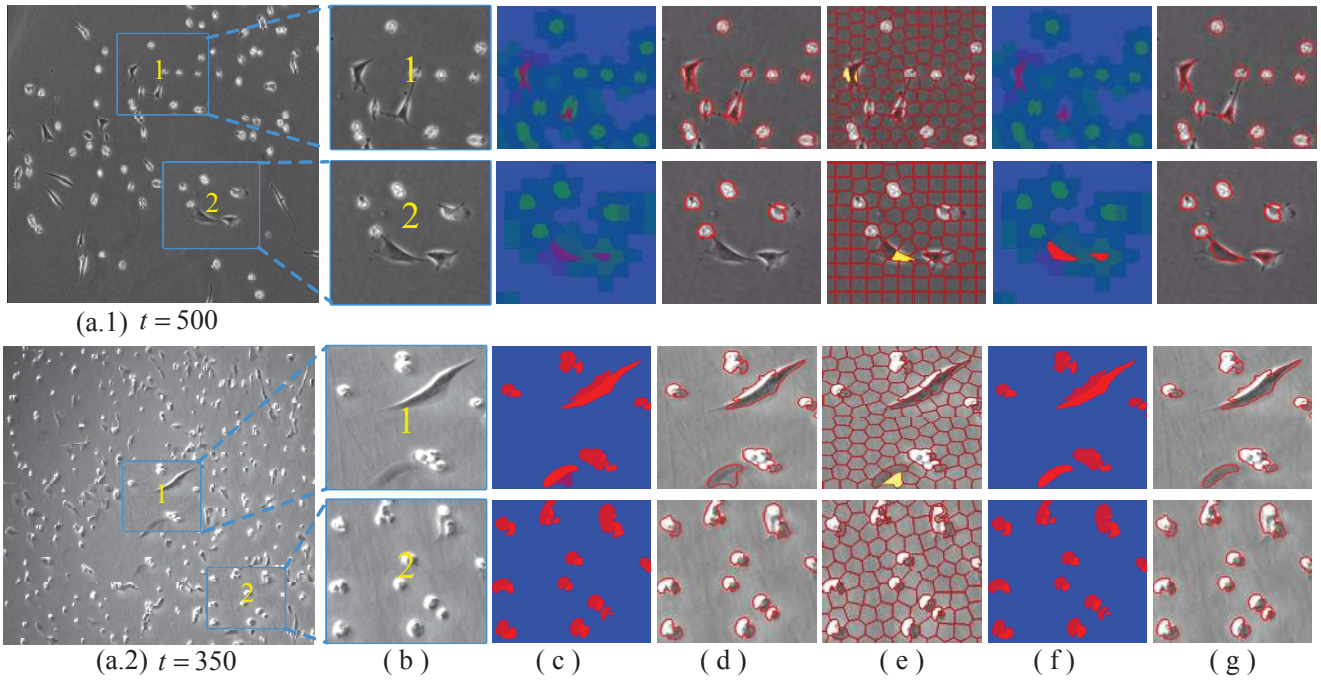


Figure 2: Examples of fixing erroneous segmentation by active sample selection and correction propagation on two images from Seq1 (top) and Seq2 (bottom). (a) Input phase contrast microscopy images; (b) Zoom-in sub-images; (c) Soft segmentation based on label propagation using annotation acquired in the first frame; (d) Hard segmentation based on the corresponding soft segmentation; (e) Actively selected atoms for correction marked by yellow color. (f) Improved soft segmentation after correction propagation; (g) Improved hard segmentation based on the improved soft segmentation.

the atoms containing errors were automatically selected by our algorithm to be verified and corrected by a human, as is shown in column (e). In column (f) and (g), we demonstrate the improved soft and hard segmentation results after human correction and its propagation. As shown, in addition to the errors of atoms selected by active atom selection, similar errors of other atoms are also effectively fixed. It is also noted that if there is no misclassification, the previous label propagation results are not influenced by subsequent human correction, as is shown in columns (d) and (g) corresponding to the sub-image (2) of Fig.2 (a.2).

The quantitative evaluation is reported in Fig. 3. As the figure shows, cell segmentation without correction (blue curves), no matter whether randomly or actively annotated samples are selected, is not comparable to the results when subsequent correction is involved (red curve). The main reason is that the visual characteristics of cells gradually change over time due to the cells' reaction to the change of its environment. If human correction is performed after the label propagation from the initial annotation, the performance is improved greatly as more information is provided by a user, since cells with different features that are not included in the first frame are identified and well segmented. In particular, correction guided by active sample selection converges more quickly than random correction, since more informative samples are selected at early stage. The sequential correction can also handle the evolving data stream but one query is selected at a time and the classifier needs to be retrained accordingly, which is apparently very slow, making the method impractical. In contrast, our method can process multiple queries at the same time thereby providing a practical solution for real-world problems. It is also observed that a user needs to correct more samples if the initial annotation is ignored, since cells in each frame within a sequence share some similar characteristics. Therefore, re-using the annotated sample reduces human efforts in correction.

5. Conclusion

In this paper, we propose an active sample selection and correction propagation algorithm for graph-based semi-supervised learning. After performing initial classification through label propagation, we actively select informative samples among unlabeled ones by minimizing the expected prediction error, and request human validation on them. Once classification error is identified and corrected, the correction is propagated to the remaining unlabeled samples through our proposed correction propagation method, which is efficient since it does not involve reconstruction of the affinity graph, resulting in effective corrections on similar errors. Experimental results demonstrate that the proposed active sample selection and correction propagation achieve high quality classification results with less human efforts for both mismatched and time-evolved data.

References

[1] *http://www.gaussianprocess.org/gpml/data/*.

[2] *https://archive.ics.uci.edu/ml/datasets/ISOLET*.

[3] *http://www.celltracking.ri.cmu.edu/downloads.html*.

[4] O. M. Aodha, N. D. F. Campbell, J. Kautz, and G. J. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2014.

[5] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291, 2004.

[6] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, Dec 2006.

[7] A. Bondu, V. Lemaire, and M. Boullé. Exploration vs. exploitation in active learning : A bayesian approach. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–7, 2010.

[8] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[9] C. Couprie, L. Grady, L. Najman, and H. Talbot. Power watershed: A unifying graph-based optimization framework. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1384–1399, July 2011.

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

[11] A. B. Goldberg, X. Zhu, A. Furger, and J. Xu. OASIS: online active semi-supervised learning. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, Aug. 2011.

[12] L. Grady. Random walks for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1768–1783, Nov. 2006.

[13] Q. Gu and J. Han. Towards active learning on graphs: An error bound minimization approach. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 882–887, 2012.

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

[15] A. Guillory and J. A. Bilmes. Label selection on graphs. In *Advances in Neural Information Processing Systems*, pages 691–699. Curran Associates, Inc., 2009.

[16] Y. Guo and R. Greiner. Optimistic active learning using mutual information. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 823–829, 2007.

[17] P. Jain and A. Kapoor. Active learning for large multi-class problems. In *Computer Vision and Pattern Recognition. IEEE Conference on*, pages 762–769, June 2009.

[18] M. Ji and J. Han. A variance minimization criterion to active learning on graphs. In N. D. Lawrence and M. A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 556–564, 2012.

[19] W. Liu and S.-F. Chang. Robust multi-class transductive learning with graphs. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 381–388, June 2009.

[20] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[21] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, October 2008.

[22] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems 20*, pages 1289–1296. 2008.

[23] H. Su, Z. Yin, S. Huh, and T. Kanade. Cell segmentation in phase contrast microscopy images via semi-supervised classification over optics-related features. *Medical Image Analysis*, 17:746–765, 2013.

[24] H. Su, Z. Yin, T. Kanade, and S. Huh. Interactive cell segmentation based on correction propagation. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 712–716, Apl. 2014.

[25] D. Tylavsky and G. R. L. Sohie. Generalization of the matrix inversion lemma. *Proceedings of the IEEE*, 74(7):1050–1052, 1986.

[26] X. Wang, B. Qian, and I. Davidson. Labels vs. pairwise constraints: A unified view of label propagation and constrained spectral clustering. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM)*, pages 1146–1151, 2012.

[27] B. Zhao, F. Wang, C. Zhang, and Y. Song. Active model selection for graph-based semi-supervised learning. In *Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference on*, pages 1881–1884, March 2008.

[28] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004.

[29] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Twentieth International Conference on Machine Learning (ICML)*, pages 912–919, 2003.

[30] X. Zhu, A. B. Goldberg, R. Brachman, and T. Dietterich. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.

[31] X. Zhu, J. Lafferty, and Z. 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.