Label Diagnosis through Self Tuning for Web Image Search

Jun Wang, Yu-Gang Jiang, and Shih-Fu Chang
Department of Electrical Engineering
Columbia University, New York, NY, USA
{jwang, yjiang, sfchang}@ee.columbia.edu

Abstract

Semi-supervised learning (SSL) relies on partial supervision information for prediction, where only a small set of samples are associated with labels. Performance of SSL is significantly degraded if the given labels are not reliable. Such problems arise in realistic applications such as web image search using noisy textual tags. This paper proposes a novel and efficient graph based SSL method with the unique capacity of pruning contradictory labels and inferring new labels through a bidirectional and alternating optimization process. The objective is to automatically identify the most suitable samples for manipulation, labeling or unlabeling, and meanwhile estimate a smooth classification function over a weighted graph. Different from other graph based SSL approaches, the proposed method employs a bivariate objective function and iteratively modifies label variables on both labeled and unlabeled samples. Starting from such a SSL setting, we present a relearning framework to improve the performance of base learner, particularly for the application of web image search. Besides the toy demonstration on artificial data, we evaluated the proposed method on Flickr image search with unreliable textual labels. Experimental results confirm the significant improvements of the method over the baseline text based search engine and the state-of-the-art SSL methods.

1. Introduction

Conventional supervised learning techniques build a mapping from observations to targets using a labeled training set. The principal assumption is that the given labels are trustable. Moreover the labeled data provide enough diversity and adequate representation of the sample space. Obviously, the supervised approaches highly rely on the quality of the training data. Besides the well-posed problem of sample selection bias [6, 10], another critical challenge is that training data may contain mislabeled samples. There has been some, but not sufficient, attention paid to this problem, such as filter based approaches for eliminating the noisy labels [3, 4, 27]. For example, in [3], ensemble classifiers were developed as filter and executed with cross validation strategy to identify and eliminate mislabeled training instances. However, all these efforts are based on the assumption of label sufficiency and require the training of supervised classifiers.

If there is partial supervision information available, i.e. a small number of training data have assigned labels, semi-supervised learning (*SSL*) approaches are commonly used to accomplish prediction and inference task. *SSL* is relevant to many real applications, when labels are usually expensive to acquire while the data acquisition is fairly cheap. Under the general framework of *SSL*, the given labels are trusted as golden truth, and the data properties, like manifold geometries are employed to carry out the inference on unlabeled data. In other words, *SSL* is often based on the principle of *trusting both label and data*. Other important considerations in *SSL* include smoothness assumption, cluster assumption, and manifold assumption [5].

Graph based *SSL* methods were recently developed with these assumptions and have shown encouraging results under agnostic settings when little prior knowledge of the data distribution and parameters is available. For example, in [23, 26], a continuous real-valued classification function is estimated thorough optimizing a predefined objective function over an undirected and weighted graph. However, previous research also shows that the performance of *SSL* methods highly relies on the quality of the given labels [22]. Therefore, label weighting is used to reduce the side effects from uninformative and noisy labels in sparse area [22]. Nevertheless, this method still can not handle wrongly labeled samples in dense regions of the data point cloud. As an intuitive demonstration, we show the mislabeling issue using the well known two-moon dataset in Figure 1, where among the eight labeled samples, two of them are falsely assigned. The results show that most of the existing techniques, either supervised or semi-supervised methods, generate erroneous prediction results (Figure 1 a-g).

In practice, the mislabeling issue occurs frequently in web image annotation due to uncontrolled labeling proce-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

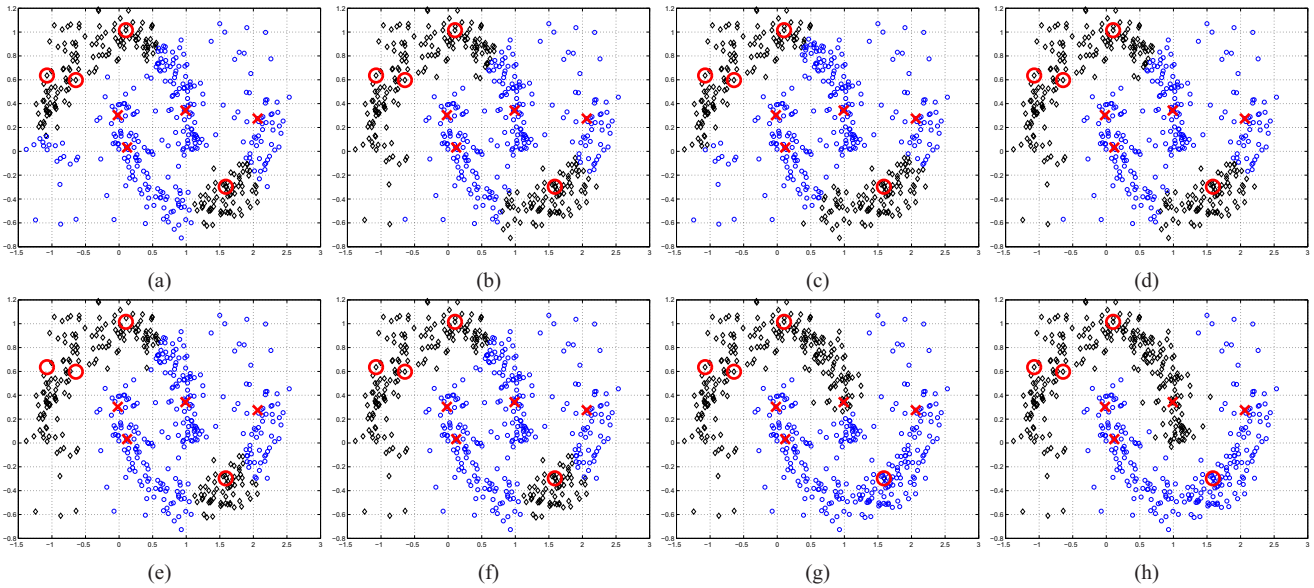

Figure 1. A demonstration of mislabeling issue on noisy two-moon data set. Large red markers indicate known labels, including wrong labels and the two-color small markers represent the classification results. a) *SVM*; b) *LapSVM* [2]; c) *RLS* [19]; d) *LapRLS* [2]; e) *GFHF* [26]; f) *LGC* [23]; g) *GTAM* [22]; h) our method *LDST*. Only *LDST* achieve fully correct results.

dure and semantic ambiguity. For example, *Flickr*, one of the most popular photo sharing website, allows users to assign textual tags when they upload images. However, it has been well recognized that there exists high inaccuracy among such manually assigned textual tags, which were observed with only around 50% accuracy [14]. Most of the current image search techniques, such as the one used in *Flickr*, utilize the textual tag associated with the images. Apparently, the error-prone tags significantly degrade the accuracy of the text search results. For instance, when the user types in the keyword "*tiger*", visually inconsistent results could be returned though all contain the key word "*tiger*" in their textual tags. Figure 2 displays the typical categories of images returned by the text search of "*tiger*", such as apex predator, butterfly, flower, tank, and golf professional. Recall the toy demonstration in Figure 1, these inaccurate tagged images can be considered as wrongly labeled samples, which may be located in either dense or sparse regions of the sample space.

Inspired by recent developments of graph based *SSL* methods, here we propose a novel method, called **L**abel **D**iagnosis through **S**elf **T**uning (*LDST*), to address this critical problem with mislabeled instances. The objective is to diagnose the quality of the given labels and remove unreliable labels while preserving visual consistency. Starting from a bivariate formulation with graph regularization, we apply a floating greedy approach to simultaneously carry out correction over labeled samples and prediction over unlabeled samples. In the case of *Flickr* image search, *LDST* is used to refine the text based image search results by systematically removing visually inconsistent images, those carry falsely assigned labels.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the related work on web image search. Section 3 presents our approach to handling mislabeling with *SSL* formulation. Section 4 provides experimental validation on both toy and the real data sets from *Flickr* image search. The conclusions and discussion are included in Section 5.

2. Related Work on Web Image Search

In response to the emerging needs in searching visual content on the web, many works on content based image retrieval (CBIR) have been proposed [15]. However, the quality of these approaches is limited because it is not easy to formulate an informative and efficient visual query. Relevance feedback is proposed to refine the user query with iterative user input [20]. However, convergence of such iterative processes is not guaranteed and the results often may not be satisfactory.

In view of the existence of abundant metadata such as captions or keywords associated with images and the maturity of text matching techniques, many works start with the text-based query and then proposed re-ranking ideas to refine the text based image search results to achieve a better query response. For example, the approaches with pseudo-relevance feedback (PRF) use initial search results as pseudo labels, among which a small number of the top-ranked samples are assumed as positive, and bottom ranked images as negative [17]. PRF methods highly rely on the

Figure 2. Typical example images using text search "*tiger*" from photo sharing website *Flickr*.

quality of the pseudo-positive/negative samples since these samples are regarded as ground truth in the subsequent learning procedure. Another category of approaches apply probabilistic models, such as constellation model [8], probabilistic latent semantic analysis and latent Dirichlet allocation [7], to train region based bag-of-words classifiers. Several limitations, including model selection and adaption, have been pointed out in [12]. The principle of information bottleneck was applied to find the optimal clusters of images that preserve the maximal mutual information [9]. However, the approximation of mutual information, which involves probability density estimation with a small number of samples, remains as a challenging problem in high dimensional image feature space. VisualRank is recently developed to exploit visual content to re-rank *Google* image search results. It uses random walk on an affinity graph to rank images based on the visual hyperlinks (similarity) among the images. Finally, the re-ranked image list is sorted based on the "importance" of graph nodes. This method has some disadvantages. First, the top ranked image sets lack diversity because visually similar images tend to share similar node importance in the graph. Second, the re-ranked images may be inconsistent if the initial text-search results have multiple dense subgraphs corresponding to different patterns. For example, the text search by "*tiger*" returns multiple disparate clusters, such as apex predator and the professional golfer, both of which have high node importance.

In summary, current web image search approaches can be categorized as supervised methods (such as PRF), probabilistic model methods, or unsupervised methods (such as information bottleneck and VisualRank). The supervised methods highly rely on the goodness of the top ranked images returned from the initial search. The unsupervised approaches are completely driven by the the data properties, like probability distribution [9] and graph geometry [12], while neglecting any initial partial supervision information contained in the top ranked images. Here, we first propose *LDST* method to handle the semi-supervised scenario with unreliable labels. Then *LDST* is applied to refine text based image search results by concurrently considering the data property and the partial supervision information obtained from the top ranked images.

3. Methodology

To handle the errors in the initial labels, we extend our previous approach *GTAM* proposed in [22] to formulate a rigorous graph transduction procedure with the capacity of handling mislabeled instances. Different from the solution of *GTAM*, here we propose a bidirectional greedy search approach to simultaneously drive wrong label correction and new label inference while preserving the smoothness and fitness of the classification function. This novel mechanism offers the unique feature to automatically prune the wrong labels to maintain a set of consistent and informative labels.

3.1. SSL with Bivariate Graph Formulation

Graph based *SSL* methods treat all samples as nodes in a graph and compute pair wise sample affinity as the estimation of edge weights. Through the connectivity among the graph nodes, the inference step on unlabeled nodes is executed via a diffusion procedure. Though there are different formulations of graph based *SSL* [25], the function estimation approaches, which approximate the graph cut solution, become popular because of the empirical success and efficiency. Most methods define a continuous classification function $f \in \mathbb{R}^{n\times c}$ (n is the number of samples and c is the number of classes) that is estimated over the graph via minimization of a cost function $\mathcal{Q}$. The cost function typically enforces a tradeoff between the smoothness of f over the weighted graph and the accuracy of fitting on the labeled nodes. Previous univariate formulations of such approaches include the Gaussian fields and harmonic functions (*GFHF*)

method [26], and the local and global consistency (*LGC*) method [23]. In both of these two methods, the objective is to derive a smooth classification function which elastically or rigidly fits on the given labels.

Due to the constraint of local fitness on labeled instances, the initially given labels dominate the above univariate system. The label inference results highly rely on the quality of the initial labels. However, in practice, the noisy and erroneous labels occur very often, like the textual tags associated with the internet images. To alleviate the dependence on the initial labels, we developed a novel bivariate formulation to drive alternating optimization on both f and binary label variable y [22]. Here we briefly describe this bivariate formation and present further revision in the following sections.

Given a weighted and undirected graph as $\mathcal{G} = \{X, E\}$, where vertex are the sample set $X = \{x_i\}, i = 1, \cdots, n$ ($n = |X|$) and the symmetric edge weight matrix $W = \{w_{ij}\}$. The sample affinity is applied to compute the edge weight through a kernel function: $w_{ij} = k(x_i, x_j)$. The node degree matrix $D = diag([d_1, \cdots, d_n])$ is defined as $d_i = \sum_{j=1}^{n} w_{ij}$. The graph Laplacian is computed as $L = D - W$ and the normalized graph Laplacian is:

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{1}$$

The binary valued variable $y \in \mathbb{B}^{n \times c}$ is set as $y_{ij} = 1$ if x_i is labeled as class j and $y_{ij} = 0$ otherwise. The objective function is then defined as:

$$\mathcal{Q}(f, y) = \frac{1}{2} \sum_{i,j} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 + \alpha \sum_i (f_i - v_i y_i)^2 \tag{2}$$

where the first part is called smoothness evaluation and the second part is fitness measurement. These two components in the above bivariate formulation are weighted by the coefficient α. The variable v_i is called label regularizer which balances the influence of labels from different classes and modulates the label importance based on the node degree. The value of v is computed as:

$$v_i = \begin{cases} p_j(x) \cdot \frac{d_i}{\sum_k y_{kj} d_k} & : \quad y_{ij} = 1 \\ 0 & : \quad otherwise \end{cases} \tag{3}$$

where $p_j(x)$ is the prior of class j and is usually set to uniform values $p_j(x) = 1/c, j = 1, \cdots, c$. If we write the label weighting term as the diagonal matrix $v = diag([v_1, \cdots, v_n])$, the quadratic objective function $\mathcal{Q}$ can be represented as the following matrix form.

$$\mathcal{Q}(f, y) = \frac{1}{2} \operatorname{tr} \left\{ f^\top \mathcal{L} f + \alpha (f - vy)^\top (f - vy) \right\} \tag{4}$$

Through minimizing the objective function, the classification function and the label matrix can be derived as:

$$(f^*, y^*) = \min_{\substack{f \in \mathbb{R}^{n \times c} \\ y \in \mathbb{B}^{n \times c}}} \mathcal{Q} \tag{5}$$

In *GTAM* [22], an alternating optimization approach is applied to iteratively minimize $\mathcal{Q}$ and estimate y. Since the minimization of objective function $\mathcal{Q}$ is convex problem with respect to variable f, the optimal f can be easily derived by zeroing the partial differential $\nabla_f \mathcal{Q}$:

$$\nabla_f \mathcal{Q} = 0 \Rightarrow f^* = (\beta \mathcal{L} + I_n)^{-1} vy = pvy \tag{6}$$

where $I_n \in \mathbb{R}^{n \times n}$ is identity matrix and $p = (\beta \mathcal{L} + I_n)^{-1}$ is called propagation matrix ($\beta = 1/\alpha$). Substitute f in the objective function $\mathcal{Q}$ by optimal f^*. Then the bivariate problem is degenerated to an univariate form as:

$$\mathcal{Q}(y) = \frac{1}{2} \operatorname{tr} \left\{ y^\top v^\top \left[p \mathcal{L} p + \beta (p - I_n)^2 \right] vy \right\} \tag{7}$$

Notice that the above problem is turned into a linearly constrained ($\sum_j y_{ij} = 1$) max cut problem and the exact solution is NP [13]. To solve this, we developed a greedy gradient based approach to gradually update y through incremental addition of labels [22].

3.2. Label Self Tuning

Compared with univariate methods, the above bivariate approach achieved much better performance due to its robustness to noisy data. The performance gain are attributed to two major factors. First the label weighting term reduces the side effect from the labels in low density and unreliable regions. Second, the iterative optimization on both variables y and f avoids prematurely committing to intractable prediction results. In order to avoid the possible state of unstable oscillation, the alternating minimization approach is only applied to unlabeled samples but not on the initial labeled set. In other works, the initially given labels are considered as golden truth and thus never changed. Because of this, *GTAM* can not handle the problem with mislabeled samples. We here revise the unilateral greedy search strategy into a bidirectional manner, which leads to our proposed approach of **L**abel **D**iagnosis through **S**elf **T**uning (*LDST*).

While preserving the optimal f, *LDST* executes floating greedy search among the most beneficial gradient directions of $\mathcal{Q}$ on both labeled and unlabeled samples. Since the label regularizer term v associated with the current label variable y, which converts the label variable into a normalized form $\tilde{y} = vy$. Following the equation 7, we derive the differential as to normalized label variable $\tilde{y}$:

$$\nabla_{\tilde{y}} \mathcal{Q} = \left[p \mathcal{L} p + \beta (p - I_n)^2 \right] \tilde{y} = \left[p \mathcal{L} p + \beta (p - I_n)^2 \right] vy \tag{8}$$

The above calculation of gradient $\nabla_{\tilde{y}}\mathcal{Q}$ measures the the change of the objective function in terms of the change of normalized label variable $\tilde{y}$. Notice that the manipulation on y is equivalent to a similar operation on $\tilde{y}$, i.e. setting $y_{ij} = 1$ leads to $\tilde{y}_{ij} = v_i$.

Since y is constraint in binary space, labeling operation is to change the value from 0 to 1 for a certain element y_{ij} in the label matrix and the unlabeling operation, i.e. removing the labels, does the reverse by setting $y_{ij} = 1 \rightarrow 0$. To reduce the value of the objective function $\mathcal{Q}$, we manipulate the label variable y in both directions, labeling and unlabeling. Note that labeling operation is carried out on the unlabeled nodes with the minimum value of the gradient $\min \nabla_{\tilde{y}}\mathcal{Q}$, while unlabeling operation is executed on the labeled nodes with the maximum value of gradient $\max \nabla_{\tilde{y}}\mathcal{Q}$. To summarize, we have the following bidirectional gradient decent search, including both labeling and unlabeling operations, to achieve the steepest reduction on the cost function $\mathcal{Q}$

$$\begin{aligned}(i^+, j^+) &= \min_{i,j} \nabla_{(vy_u)}\mathcal{Q}; \quad y_{i^+j^+} = 1 \\ (i^-, j^-) &= \max_{i,j} \nabla_{(vy_l)}\mathcal{Q}; \quad y_{i^-j^-} = 0 \end{aligned} \tag{9}$$

where (i^+, j^+) and (i^-, j^-) are the optimal elements of variable y for labeling and unlabeling operations, respectively. Different from the labeling procedure, the optimal element for unlabeling operation is only investigated on the positions of variable y_l where the element has the nonzero values. In other words, through each bidirectional gradient decent iteration, we add one most reliable label, and meanwhile remove one least confident label. Since the label regularizer term v is associated with the current labels as shown in Equation 3, we need to update v after each individual operation, either labeling or unlabeling.

3.3. Final Algorithm

Here we finalize the *LDST* method in chart 1. From this chart, in the first s iterations, a number of unlabeling and labeling operations are executed in order to eliminate the problematic labels and add trustable new labels. We refer to this stage as *LDST-self-tuning*. In this self tuning stage, one new label is added to the labeled set after one unreliable label is eliminated to maintain a fix number of labels. Moreover, each individual operation of labeling and unlabeling leads to the update of label regularization matrix v. After executing certain steps of label self tuning, the subsequent stage, called *LDST-propagation*, is conducted to propagate the labels to unlabeled set. Theoretically, the algorithm stops when all the unlabled samples are labeled. However, this may result in prohibitive computation if the data size is huge. There are two strategies to speed up the algorithm to meet the computational needs in realistic application. First, the iterative procedure can be early terminated after obtaining enough labels. The final prediction results are computed using the propagation Equation 6. Second, the computation associated with matrix multiplication for calculating gradient $\nabla\mathcal{Q}^{t+1}_{(vy_u)}$ can be converted to vector addition since each step only involves the change of a single vector entry in $\nabla\mathcal{Q}^{t}_{(vy_u)}$, similar to the incremental labeling method reported in [21].

Input: data set $\mathcal{X} = \{\mathcal{X}_l, \mathcal{X}_u\}$, the graph $\mathcal{G}\{X, E\}$ and the corresponding constants:
normalized graph Laplacian $\mathcal{L}$;
propagation matrix p;
node degree matrix D;
gradient constant $g = p\mathcal{L}p + \beta(p - I_n)^2$;
initial label variable y^0;
label regularizer v^0.
Output: optimal prediction function f^* and labels y^*.

iteration counter $t = 0$;
self tuning iteration number s;
while $\mathcal{X}_u \neq \emptyset$ **do**
  compute gradient $\nabla\mathcal{Q}^t_{(vy_u)} = gv^t y^t$;
  if $t \leq s$ **then**
    $(i^-, j^-) = \max_{i,j} \nabla\mathcal{Q}^t_{(vy_l)}$;
    $y_{i^-,j^-} = 0$;
    update $\mathcal{X}_l, \mathcal{X}_u$;
    recalculate v^t;
  end
  $(i^+, j^+) = \min_{i,j} \nabla\mathcal{Q}^t_{(vy_u)}$;
  $y_{i^+,j^+} = 1$;
  update $\mathcal{X}_l, \mathcal{X}_u$;
  $t = t + 1$;
  recalculate v^t;
end
return $y^*, f^* = pvy$.

Algorithm 1: The algorithmic chart of label diagnosis through self tuning approach.

4. Experiments

4.1. Toy Demonstration

We first use synthetic noisy two-moon artificial data for intuitive demonstration. We manipulated the toy data from [2] by adding 100 random noisy instances to obtain a non-separable point cloud containing 500 2D samples, as shown in Figure 1. Larger red markers are the labeled samples and the shape represents different classes, positive or negative. Each class is assigned four labeled samples, among which one is mislabeled.

For this challenging classification task with wrong labels, we compare different approaches, including supervised, like standard *SVM*, and graph based *SSL* algorithms,

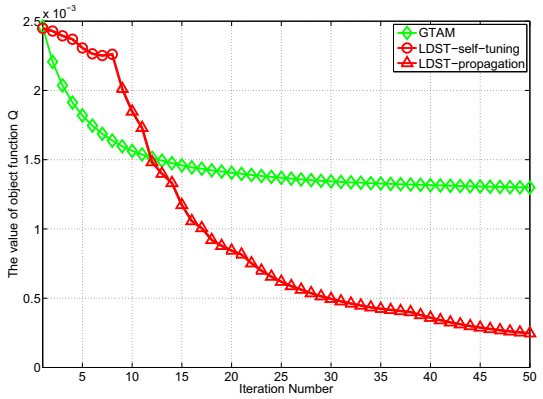

Figure 3. The values of the cost function $\mathcal{Q}$ during optimization procedure of *LDST* and *GTAM* methods.

such as *GFHF* [26], *LGC* [23], and *GTAM* [22]. Moreover, previous work shows that *LapSVM* and *LapRLS* outperform some existing *SSL* methods [2]. Therefore, we also include these in the comparison study. For all the graph based approaches, we use the common setting to build a KNN graph with the same number of neighborhoods ($k = 6$) and adaptive RBF kernel size [21]. For other parameters, we adopt the best setting reported in previous literatures.

The classification results by different methods are shown in Figure 1 (a)-(h). From this demonstration, the following findings can be obtained. First, the performance of supervised methods is heavily degraded due to the blind trust on the false labels and the exclusive reliance on the labeled data. Second, most *SSL* methods generate erroneous results even both the labeled and unlabeled data are considered. The reason lies in the fact that the labeled samples outweigh the unlabeled data in driving the inference process. Especially, some algorithms, like *GFHF*, clamp the prediction results on given labels. *LGC* incorporates elastic fitness term but can not rectify the wrong labels. *GTAM* achieved the best accuracy among the existing methods due to its bivariate formulation and iterative label propagation procedure. However, it still lost the manifold separation due to the high level of noise influence. Thus we conclude that existing *SSL* methods are incapable of identifying and eliminating the mislabeled samples. On the contrary, the proposed *LDST* uses the self tuning stage to eliminate unreliable labels leading to accurate label prediction results without breaking the structure of the two data clusters, as shown in Figure 1 (h). Furthermore, since *GTAM* achieved high accuracy close to *LDST*, and both share the common technique of gradient decent, we analyze the cost function value $\mathcal{Q}$ during the optimization procedure of these two methods by showing the first 50 iterations in Figure 3. This figure clearly shows that after pruning the wrong labels by self

Method	SVM	LapSVM	RLS	LapRLS	GFHF	LGC	GTAM	LDST
Error (%)	34.64	30.16	34.01	30.26	38.76	23.77	5.99	**0.91**
Std (%)	7.03	10.63	11.23	10.67	3.69	6.82	11.24	**2.43**

Table 1. The mean and standard deviation of the error rate on 20 random runs of the toy experiment.

tuning (first 8 iterations marked as red circle) to get a consistent label set, the cost rapidly descends afterwards in the propagation stage. This observation demonstrates that most of the prediction accuracy can be attributed to the label self tuning procedure.

In addition, a comprehensive comparison study was conduct by 20 rounds of random tests. For each round, three correct labels and one wrong label are randomly assigned to each class. The same graph and parameters, like the self tuning iteration number $s = 8$, are fixed for all the runs. The mean and standard deviation of the classification errors are recorded in Table 1. From this statistical evaluation, *LDST* archives much higher and more stable performance under the situation of mislabeled samples.

4.2. Web Image Search

There is abundant textual tag information available in most of the current image websites. It makes sense to exploit such textual tags in image search. To address the errors associated with the imperfect tags, we use the proposed *LDST* method to greatly improve the text based image search results. Two assumptions are made. First, the desired targets are at least one of the majority patterns in the initially returned image set. Second, the top-ranked images are more likely to include the targets. These two assumptions are typically valid in the practical situations. Based on these, the top-ranked image are first truncated to create a set of pseudo positive labels, while the images with lower ranking orders are treated as unlabeled samples. Then *LDST* is applied to tune the imperfect labels and further refine the rank list.

To evaluate our approach on the web image search task, a total of nine categories of images are acquired from the photo sharing website *Flickr* using text search. The selected categories cover a diverse range of targets, including animals, plants, man-made objects and scenes. For each set of text search results, about 1500 returned images are collected for re-ranking. Example images corresponding to these text queries are shown in Figure 4.

For image feature representation, we adopt the widely used Bag-of-Visual-Words (*BoW*) derived from local key points, which has shown effective in many applications of object and scene classification. We use difference of Gaussian as key point detector and SIFT as descriptor [16]. To quantize the local features to visual words, we adopt the soft-assignment strategy which has been shown effective in

Figure 4. Example images of text search results from *flickr.com*. A total of nine text queries are used: *dog*, *tiger*, *panda*, *bird*, *flower*, *airplane*, *forbidden city*, *statue of liberty*, *golden bridge*.

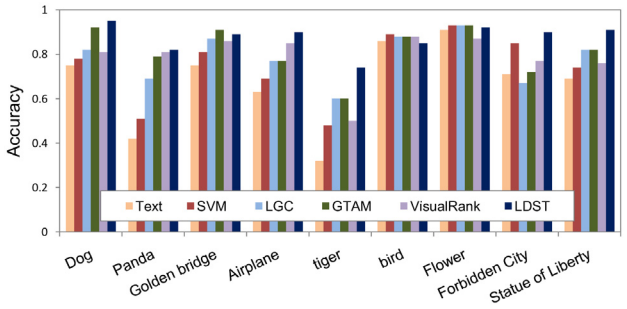

Figure 5. Comparison of the precision of the top 100 ranked images over different categories of images.

[11, 18].

The pair wise affinity value is computed using cosine similarity between *BoW* vectors. The number of nearest neighbors is uniformly set as 200 (a typical setting for cosine similarity graph [1, 22]) to construct KNN graphs for the returned images from each individual query. Since there is no clear cue for selecting negative samples for each individual query, the classification task is degenerated to a ranking problem [24]. Here the top ranked 60 samples are treated as pseudo positive labels. Label self tuning is used to remove visually inconsistent samples and afterward propagation is done to rank the remaining images (the number of self tuning iteration s is uniformly set as 30). We compare with other automatic re-ranking methods, like Pseudo-relevance feedback (*PRF*) framework [17]. Specifically, we designed *PRF-SVM*, *PRF-LGC* and *PRF-GTAM* for comparison. In addition, we compare with the recent VisualRank technique [12], which has shown empirical success on product image search on *Google*. The same parameter setting is applied as suggested by [12]. Moreover, all the graph based approaches use the same graph, as described earlier. The precision of the top 100 ranked images are calculated to evaluate the performance, as shown in Figure 5.

From Figure 5, *LDST* achieves significant performance

Method	Text	SVM	LGC	GTAM	VisualRank	LDST
Accuracy (%)	67.11	74.22	79.89	81.44	79.00	87.56

Table 2. The accuracy of the top ranked *Flickr* images by different approaches.

improvement over most semantic categories, like *tiger*, *panda*, and *dog*. For these cases, the visual content of targets exhibits consistent pattern though there is strong ambiguity associated with the keywords. The performance of VisualRank is degraded since the returned image set from text search contains multiple conflicting patterns. However, *LDST* does not show much gain on the categories of *bird* and *flower*. Because the visual content of the positive samples associated with these two tags is quite diverse, which affects the stability of the label self tuning procedure. Overall, *LDST* improves the average accuracy of text search results from 67.11% to 87.56% on the nine categories (Table 2). Compared to the state-of-the-art methods, the *LDST* approach also enjoys a clear performance gain (7.5% over *GTAM*, and 10.8% over VisualRank).

There are two parameters in *LDST* for web image search, the initial number of positive l and the number of tuning iteration s. In the above experiments, we empirically set $l = 60$ images from top ranked list as positive samples and fix $s = l/2$. We have carried out an extensive study by varying the value of l from 20 to 100. The result shows that *LDST* achieved fairly consistent and stable performance under different choices of l with the precision of the top 100 ranked images as $87.23 \pm 0.6\%$. In addition, the proposed method can be made very efficient by applying the superposable update model developed in [21]. The current implementation takes only a few seconds to rerank more than 1500 images for each query on a regular PC.

5. Conclusion Remarks

The main contributions of this article consist of the bivariate formulation of graph-based semi-supervised learning (*SSL*) for handling errors in the initial labeled set, and its application for re-ranking text-based web image search results. Conventional *SSL* methods fail in such cases due to their trust on initial unreliable labels as golden truth. In this paper, in order to identify and eliminate wrong labels, we propose a novel approach, named **L**abel **D**iagnosis though **S**elf **T**uning (*LDST*). It combines label diagnosis, manipulation, and propagation in a uniform optimization framework. Specifically, a floating gradient greedy search method is applied to manipulate the most beneficial samples and optimize the bivariate objective function.

We validate the effectiveness of the proposed *LDST* method through extensive experiments with artificial data set and text based web image search. The textual tags are treated as potentially incorrect labels and the *LDST* method

is applied to correct label mistakes and propagate label information over the entire collection. The experimental results over nine diverse categories of *Flickr* images confirm the significant performance gain over both text search baseline and the state-of-the-art reranking methods.

The proposed *LDST* method is general in the sense that no prior training process is required. Therefore, it is readily applicable to broad search scenarios on the Internet using existing search engines.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. *Proc. AISTAT*, 2005.

[2] 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, 2006.

[3] C. E. Brodley and M. A. Friedl. Identifying and eliminating mislabeled training instances. In *Proc. AAAI*, 1996.

[4] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11, 1999.

[5] O. Chapelle, B. Schölkopf, A. Zien, and I. NetLibrary. *Semi-supervised learning*. MIT Press, 2006.

[6] M. Dudık and P. S. J. Schapire, R. E. Correcting sample selection bias in maximum entropy density estimation. *Proc. NIPS*, 17, 2005.

[7] R. Fergus, F. F. Li, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proc. ICCV*, 2005.

[8] R. Fergus, P. Perona, and A. Zisserman. A Visual Category Filter for Google Images. In *Proc. ECCV*, 2004.

[9] W. H. Hsu, L. S. Kennedy, and S. F. Chang. Video search reranking via information bottleneck principle. In *Proc. ACM Multimedia*, 2006.

[10] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf. Correcting Sample Selection Bias by Unlabeled Data. *Proc. NIPS*, 19, 2007.

[11] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proc. CIVR*, 2007.

[12] Y. Jing and S. Baluja. VisualRank: Applying PageRank to Large-Scale Image Search. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12, 2008.

[13] R. M. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, 43:85–103, 1972.

[14] L. S. Kennedy, S. F. Chang, and I. V. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proc. ACM MIR*, 2006.

[15] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[16] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[17] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proc. ACM MM*, 2007.

[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. *Proc. CVPR*, 2008.

[19] R. Rifkin and R. Lippert. Notes on regularized least squares. Technical Report MIT-CSAIL-TR-2007-025, Computer Sciences and Artificial Intelligence Laboratory, MIT, 2007.

[20] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[21] J. Wang, S. F. Chang, X. Zhou, and S. T. C. Wong. Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels. In *Proc. CVPR*, 2008.

[22] J. Wang, T. Jebara, and S. F. Chang. Graph transduction via alternating minimization. In *Proc. ICML*, 2008.

[23] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Proc. NIPS*, volume 16, 2004.

[24] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf. Ranking on Data Manifolds. In *Proc. NIPS*, volume 16, 2004.

[25] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, CS, University of Wisconsin-Madison, 2005.

[26] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. ICML*, 2003.

[27] X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *Proc. ICML*, 2003.