Learning Instance Specific Distance for Multi-Instance Classification

Hua Wang, Feiping Nie, Heng Huang
Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, feipingnie@gmail.com, heng@uta.edu

Abstract

Multi-Instance Learning (MIL) deals with problems where each training example is a bag, and each bag contains a set of instances. Multi-instance representation is useful in many real world applications, because it is able to capture more structural information than traditional flat single-instance representation. However, it also brings new challenges. Specifically, the distance between data objects in MIL is a set-to-set distance, which is harder to estimate than vector distances used in single-instance data. Moreover, because in MIL labels are assigned to bags instead of instances, although a bag belongs to a class, some, or even most, of its instances may not be truly related to the class. In order to address these difficulties, in this paper we propose a novel Instance Specific Distance (ISD) method for MIL, which computes the Class-to-Bag (C2B) distance by further considering the relevances of training instances with respect to their labeled classes. Taking into account the outliers caused by the weak label association in MIL, we learn ISD by solving an ℓ_{0+}-norm minimization problem. An efficient algorithm to solve the optimization problem is presented, together with the rigorous proof of its convergence. The promising results on five benchmark multi-instance data sets and two real world multi-instance applications validate the effectiveness of the proposed method.

Introduction

Multi-Instance Learning (MIL) (Dietterich, Lathrop, and Lozano-Pérez 1997) is a new paradigm in machine learning that addresses the classification of bags. In MIL, each *bag* is a collection of *instances* with features associated to the instance. The aim of MIL is to infer bag level labels based on the assumption that a positive bag contains at least one positive instance, whereas a negative bag contains negative instances only. MIL has been found useful in a number of real world applications (Maron and Ratan 1998; Zhang and Goldman 2002; Zhou and Zhang 2007; Zhou, Sun, and Li 2009; Wang, Hu, and Chia 2010; Li et al. 2011).

A prominent advantage of MIL lies in the fact that many real objects have inherent structures, and by adopting the multi-instance representation we are able to represent such objects more naturally and capture more information than simply using the flat single-instance representation. For example, suppose we can partition an image into several parts. In contrast to representing the whole image as a single-instance, if we represent each part as an instance, the partition information is captured by the multi-instance representation; and if the partition is meaningful (*e.g.*, each part corresponds to a region of saliency), the additional information captured by the multi-instance representation may be helpful to make the learning task easier to deal with.

Multi-instance representation, though usually useful, also brings new challenges for statistical learning. First, because in MIL an object is represented as a bag of instances, the distance between objects turns out to be a set-to-set distance. Thus, compared to single-instance data using vector distance such as Euclidian distance, distance estimation in MIL is more complicated. Second, in MIL labels are assigned to bags but not instances, which is often called as "weak label association". As a result, although a bag belongs to a class, some, or even most, of its instances may not be truly related to the class. For example, region (instance) A of the top left training image in Figure 1 only characterizes class "ship", while the entire image is labeled with both "ship" and "person'. Intuitively, instance A should have much less, or even no, impact when predicting label "person" for a query image, whereas contribute a lot when predicting label "ship". With these recognitions, in this work we explore the difficulties, as well as opportunities, of MIL to improve the classification performance on multi-instance data.

Instance Specific Distance (ISD) for MIL

Because traditional *Bag-to-Bag (B2B)* distance often does not truly reflect the class relationships between data objects (Boiman, Shechtman, and Irani 2008), in this paper we consider to directly assess the relevance between classes and query objects, and propose a novel *Class-to-Bag (C2B)* distance for MIL. Specifically, as illustrated in Figure 1, we consider each class as a "super-bag", which comprises all instances from the bags that belong to this class. The elementary distance from an instance in a super-bag to a query bag (red and blue arrows) is first estimated, then the C2B distance from the class to the query object is computed as the sum of the elementary distances from all the instances in the super-bag to the query bag. Furthermore, we also consider

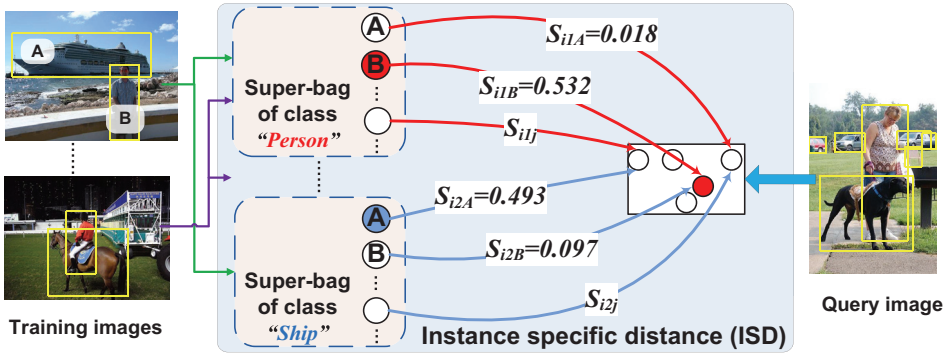

Figure 1: An illustration of the proposed Instance Specific Distance (ISD), which is defined as the Class-to-Bag (C2B) distance from a super-bag corresponding to a class to a query bag. Instead of considering every instance in a super-bag equally when estimating the C2B distance, we learn a Significance Coefficient (SC) for each instance with respect to each of its labeled class, denoted as s_{ikj}, to reflect its relative importance when predicting labels for a query object. For example, the learned SC of instance A with respect to class "person" is 0.018 indicating that it has small impact in predicting label "person"; while its learned SC with respect to class "ship" is 0.493 indicating that it is considerably important when predicting label "ship".

the relative importance of a training instance with respect to its labeled classes by assigning it one weight for each labeled class, called as *Significance Coefficient (SC)* . Ideally, the learned SC of an instance with respect to its true belonging class should be large, whereas its SC with respect other classes should be small. We call the learned C2B distance as *Instance Specific Distance (ISD)* , which is interesting from a number of perspectives as following.

- Through the proposed ISD, we embrace, rather than ignore, the complexity of multi-distance data. To the best of our knowledge, we are the first to *explicitly* address the weak label association in MIL.
- The learned SCs of an instance reflect its relative importance with respect to its labeled classes, thereby provide a clearer insight of a multi-instance data set.
- Different from traditional B2B distance, the learned C2B distance can be directly used to predict object labels.
- In order to address the outliers caused by the weak label association in MIL, we learn ISD by solving an ℓ_{0+}-norm minimization problem. An novel yet efficient algorithm to solve the optimization problem is presented, and its convergence is rigorously proved.
- Promising experimental results on five benchmark multi-instance data sets and two real world multi-instance applications show that our method is highly competitive to state-of-the-art MIL methods.

Problem Formalization

Given a multi-instance data set with K classes and N training samples $\mathcal{D} = \{(X_i, \mathbf{y}_i)\}_{i=1}^{N}$, each $X_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i,n_i}] \in \mathbb{R}^{d \times n_i}$ is a bag of n_i instances, where $\mathbf{x}_{ij} \in \mathbb{R}^d$ is an instance. The label indicator $\mathbf{y}_i \in \{0,1\}^K$ is a binary vector. In the setting of MIL, if there exists $g \in \{1, \ldots, n_i\}$ such that $\mathbf{x}_{ig}$ belongs to the k-th class, X_i is assigned to the k-th class and $\mathbf{y}_i(k) = 1$; otherwise $\mathbf{y}_i(k) = 0$. Yet the concrete value of the index g is unknown. We write $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_K]^T$. If $\sum_{k=1}^{K} Y_{ik} = 1$, *i.e.*, each data object (bag) belongs to exactly one class, the data set is a single-label data set; if $\sum_{k=1}^{K} Y_{ik} \geq 1$, *i.e.*, each bag may be associated with more than one class label, the data set is a multi-label data set (Wang, Ding, and Huang 2010a; 2010b). Our task is to learn from $\mathcal{D}$ a classifier that is able to predict labels for unseen bags.

Instance Specific Distance (ISD) for MIL

In this section, we first propose a novel ISD for MIL and develop the optimization objective to learn it. Then we present a novel yet efficient algorithm to solve the optimization problem, followed by the rigorous proof of its convergence.

Formulation of ISD

In order to compute the C2B distance, we represent every class as a super-bag consisting of the instances from all its training bags:

$$C_k = \{\mathbf{x}_{ij} \mid i \in \pi_k\}, \tag{1}$$

where $\pi_k = \{i \mid Y_{ik} = 1\}$ is the index set of all training bags belonging to the k-th class. Then we may compute the elementary distance from an instance of a super-bag C_k to a data object bag $X_{i'}$ using the distance from the instance to its nearest neighbor instance in $X_{i'}$:

$$d_k(\mathbf{x}_{ij}, X_{i'}) = \|\mathbf{x}_{ij} - \tilde{\mathbf{x}}_{ij}\|^2, \quad \forall\, i \in \pi_k, \tag{2}$$

where $\tilde{\mathbf{x}}_{ij}$ is the nearest neighbor of $\mathbf{x}_{ij}$ in $X_{i'}$. Hence the C2B distance from a super-bag C_k to $X_{i'}$ is computed as:

$$D(C_k, X_{i'}) = \sum_{i \in \pi_k} \sum_{j=1}^{n_i} d_k(\mathbf{x}_{ij}, X_{i'}) \quad . \tag{3}$$

The C2B distance defined in Eq. (3) does not take into account the instance level labeling ambiguity caused by the weak label association, thus we further develop it by weighting the instances in a super-bag upon their relevance.

Due to the weak association between instances and labels, not all the instances in a super-bag really characterize the corresponding class. For example, in Figure 1 instance A

(ship) is in the super-bag of "person" class, because the entire image (top left) is labeled with both "person" and "ship". As a result, intuitively, we should give it a smaller, or even no, weight when determining whether to assign "person" label to a query image; and give it a higher weight when deciding "ship" label. More precisely, let s_{ikj} be the weight associated with $\mathbf{x}_{ij}$ with respect to the k-th class, we wish to learn the C2B distance from the k-th class to $X_{i'}$ as:

$$D\left(C_k, X_{i'}\right) = \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ikj}\; d_k\left(\mathbf{x}_{ij}, X_{i'}\right) \quad . \tag{4}$$

We call s_{ikj} as the *Significance Coefficient (SC)* of $\mathbf{x}_{ij}$ with respect to the k-th class, and the C2B distance computed in Eq. (4) as our proposed Instance Specific Distance (ISD).

Optimization Objective

Armed with the C2B distance defined in Eq. (4), we can learn s_{ikj} for instance $\mathbf{x}_{ij}$ with respect to the k-th class using least square regression to minimize the following objective:

$$\begin{aligned} J_1\left(\mathbf{w}_k, \mathbf{s}_{ik}\right) &= \sum_{i \in \pi_k} \left(\mathbf{w}_k^T \sum_{j=1}^{n_i} \mathbf{x}_{ij} s_{ikj} - Y_{ik} \right)^2 + \gamma \|\mathbf{w}_k\|^2, \\ &= \sum_{i \in \pi_k} \left(\mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik} \right)^2 + \gamma \|\mathbf{w}_k\|^2, \qquad (5) \\ s.t. \quad & \mathbf{s}_{ik} \geq 0,\ \mathbf{e}^T \mathbf{s}_{ik} = 1, \quad \forall\, 1 \leq k \leq K, \end{aligned}$$

where $\mathbf{w}_k \in \mathbb{R}^d$ is the projection vector for the k-th class, $\mathbf{s}_{ik} = \left[s_{ik1}, \ldots, s_{ikn_i}\right]^T \in \mathbb{R}^{n_i}$, $\gamma > 0$ is the regularization parameter to avoid over-fitting, and $\mathbf{e} = [1, \ldots, 1]^T$ is a constant vector with all entries to be 1. Note that, instead of learning one single regression vector for all K classes, we decide to learn K different projection vectors $\mathbf{w}_k\,(1 \leq k \leq K)$, one for each class, to capture the class-wise data variances following (Xiang, Nie, and Zhang 2008; Wang, Hu, and Chia 2010). In addition, we constrain the overall weight of a single bag to be unit, *i.e.*, $\mathbf{s}_{ik} \geq 0,\ \mathbf{e}^T \mathbf{s}_{ik} = 1$, such that all the training bags are fairly used. This constraint is equivalent to require the ℓ_1-norm of $\mathbf{s}_{ik}$ to be 1 and implicitly enforce sparsity on $\mathbf{s}_{ik}$ (Tibshirani 1996; Nie et al. 2010), which is in accordance with the fact that one class label of a bag usually arises from only one or a few of its instances but not all.

Despite its clear intuition and closed-form solution, the least square loss function used in Eq. (5) is sensitive to outliers. Due to the weak label association, outlier instances are abundant in multi-instance data, such as instance A in the super-bag "person" and instance B in the super-bag "ship" in Figure 1. Therefore, a robust loss function is expected:

$$\begin{aligned} J_2\left(\mathbf{w}_k, \mathbf{s}_{ik}\right) &= \sum_{i \in \pi_k} \left| \mathbf{w}_k^T \sum_{j=1}^{n_i} \mathbf{x}_{ij} s_{ikj} - Y_{ik} \right|^p + \gamma \|\mathbf{w}_k\|^2, \\ &= \sum_{i \in \pi_k} \left| \mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik} \right|^p + \gamma \|\mathbf{w}_k\|^2, \qquad (6) \\ s.t. \quad & \mathbf{s}_{ik} \geq 0,\ \mathbf{e}^T \mathbf{s}_{ik} = 1, \quad \forall\, 1 \leq k \leq K \quad . \end{aligned}$$

When $p = 2$, J_2 is exactly J_1. When $0 < p < 2$, the outliers have less importance in the first term of J_2 than the sum of squared residues in J_1 (Nie et al. 2010). The smaller p is, the more robust against outliers J_2 is. When $p = 1$, the loss function is the sum of the absolute values of the residues, which is similar to LASSO (Tibshirani 1996). In this paper, we set $p \to 0$, which is more desirable for MIL. Therefore, we call solving objective J_2 as ℓ_{0^+}*-norm minimization* method.

Optimization Algorithm and Its Convergence

When $p = 1$, the optimization problem J_2 can be reformulated and solved as a LASSO (Tibshirani 1996) problem. However, when $p \to 0$ as expected in MIL, the optimization problem is hard to solve in general and traditional optimization methods can not be used. In the rest of the section, we derive a novel yet efficient optimization algorithm to solve J_2 and present the rigorous proof of its convergence.

Following standard optimization procedures, we alternatively optimize the two variables, $\mathbf{w}_k$ and $\mathbf{s}_{ik}$, of J_2.

When $\mathbf{w}_k$ is fixed, the objective in Eq. (6) can be decoupled by different i. Then we can solve the following simpler problem for each i independently:

$$\min_{\mathbf{s}_{ik} \geq 0,\ \mathbf{e}^T \mathbf{s}_{ik} = 1} \left| \mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik} \right|^p, \tag{7}$$

which is equivalent to the following Quadratic Programming(QP) problem:

$$\min_{\mathbf{s}_{ik} \geq 0,\ \mathbf{e}^T \mathbf{s}_{ik} = 1} \left(\mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik} \right)^2. \tag{8}$$

Note that the number of instances in a bag, *i.e.*, n_i, is usually small, thus the optimal solution to this QP problem can be easily obtained without heavy computation.

When fixing $\mathbf{s}_{ik}$, we propose an efficient algorithm to minimize the objective in Eq. (6) with respect to $\mathbf{w}_k$. Denote $X_{(k)} = [X_1 \mathbf{s}_{1k}, X_2 \mathbf{s}_{2k}, \ldots, X_i \mathbf{s}_{ik}, \ldots, X_{N_k} \mathbf{s}_{N_k k}] \in \mathbb{R}^{d \times N_k}\ (i \in \pi_k)$ where $N_k = |\pi_k|$ is the number of bags belonging to the k-th class, and $y_{(k)} = [Y_{1k}, Y_{2k}, ..., Y_{Nk}]^T \in \mathbb{R}^{N_k}$. The detailed algorithm to minimize J_2 in Eq. (6) is described in Algorithm 1.

Now, we will prove that Algorithm 1 monotonically decreases the objective value of J_2 in Eq. (6) in each iteration by the following lemma and theorem, which guarantees the convergence of the algorithm.

Lemma 1 $\left|\tilde{x}\right|^p - \frac{p}{2}\left|x\right|^{p-2}\tilde{x}^2 \leq \left|x\right|^p - \frac{p}{2}\left|x\right|^{p-2}x^2$

Proof: Denote $\sigma = \left|\frac{\tilde{x}}{x}\right|^2$ and $h(\sigma) = 2\sigma^{\frac{p}{2}} - p\sigma + p - 2$, then $h'(\sigma) = p\sigma^{\frac{p-2}{2}} - p$ and $h''(\sigma) = \frac{p(p-2)}{2}\sigma^{\frac{p-4}{2}}$. Obviously, when $0 < p \leq 2$, we have $h''(\sigma) \leq 0$, then $h(\sigma)$ is a concave function. Note that $h'(1) = 0$, so $h(1) = 0$ is a global maximum of $h(\sigma)$, that is, $h(\sigma) \leq 0$. Therefore, $\left|x\right|^p \left(2\left|\frac{\tilde{x}}{x}\right|^p - p\left|\frac{\tilde{x}}{x}\right|^2 + p - 2\right) \leq 0$. Thus we have

$$\begin{aligned} & \left|x\right|^p \left(2\left|\frac{\tilde{x}}{x}\right|^p - p\left|\frac{\tilde{x}}{x}\right|^2 + p - 2\right) \leq 0 \\ \Rightarrow\ & \left|\tilde{x}\right|^p - \tfrac{p}{2}\left|x\right|^{p-2}\tilde{x}^2 \leq \tfrac{2-p}{2}\left|x\right|^p \\ \Rightarrow\ & \left|\tilde{x}\right|^p - \tfrac{p}{2}\left|x\right|^{p-2}\tilde{x}^2 \leq \left|x\right|^p - \tfrac{p}{2}\left|x\right|^{p-2}x^2, \end{aligned} \tag{9}$$

which complete the convergence proof. ■

Data: $X_i(1 \leq i \leq N) \in \mathbb{R}^{d \times n_i}$, $Y \in \mathbb{R}^{N \times K}$
Result: $\mathbf{w}_k(1 \leq k \leq K) \in \mathbb{R}^d$,
$\mathbf{s}_{ik}(1 \leq i \leq N, 1 \leq k \leq K) \in \mathbb{R}^{n_i}$
Initialize $D_k(1 \leq k \leq K) \in \mathbb{R}^{N \times N}$ as an identity matrix;
repeat
1. Update $\mathbf{w}_k(1 \leq k \leq K)$ by $\mathbf{w}_k = (X_{(k)} D_k X_{(k)}^T + \gamma I)^{-1} X_{(k)} D_k y_{(k)}$;
2. Update $\mathbf{s}_{ik}(1 \leq i \leq N, 1 \leq k \leq K)$ by $\min_{\mathbf{s}_{ik} \geq 0,\ \mathbf{e}^T \mathbf{s}_{ik} = 1} \left(\mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik}\right)^2$;
3. Calculate the diagonal matrix $D_k(1 \leq k \leq K)$, where the i-th diagonal element is $\frac{p}{2} \left|\mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik}\right|^{p-2}$;
until *Converges*;

Algorithm 1: An efficient iterative algorithm to minimize the objective in Eq. (6).

Theorem 1 *Algorithm 1 will monotonically decrease the objective of Eq. (6) in each iteration.*

Proof: Let $a_{ik} = \mathbf{w}_k^T X_i \mathbf{s}_{ik} - Y_{ik}$. Denote the updated $\mathbf{w}_k$ by $\tilde{\mathbf{w}}_k$, and the corresponding updated a_{ik} by $\tilde{a}_{ik}$. According to step 1 in the algorithm, we have

$$\sum_{i=1}^{n} \frac{p}{2} |a_{ik}|^{p-2} \tilde{a}_{ik}^2 + \gamma \|\tilde{\mathbf{w}}_k\|^2 \leq \sum_{i=1}^{n} \frac{p}{2} |a_{ik}|^{p-2} a_{ik}^2 + \gamma \|\mathbf{w}_k\|^2 \tag{10}$$

According to Lemma 1, we have

$$\sum_{i=1}^{n} |\tilde{a}_{ik}|^p - \sum_{i=1}^{n} \frac{p}{2} |a_{ik}|^{p-2} \tilde{a}_{ik}^2 \leq \sum_{i=1}^{n} |a_{ik}|^p - \sum_{i=1}^{n} \frac{p}{2} |a_{ik}|^{p-2} a_{ik}^2 \tag{11}$$

Adding Eq. (10) and Eq. (11) in both sides, we have

$$\sum_{i=1}^{n} |\tilde{a}_{ik}|^p + \gamma \|\tilde{\mathbf{w}}_k\|^2 \leq \sum_{i=1}^{n} |a_{ik}|^p + \gamma \|\mathbf{w}_k\|^2 \tag{12}$$

Thus the objective value of J_2 is decreased. According to step 2 in the algorithm, the objective value of J_2 is further decreased, which completes the proof of the theorem. ■

Note that the objective J_2 in Eq. (6) is obviously lower bounded by 0, thus Algorithm 1 will converge.

Incorporating Class-Specific Adaptations

Besides $\mathbf{s}_{ki}$, the objectives J_1 and J_2 have another variable $\mathbf{w}_k$. Upon solution, the learned $\mathbf{w}_k$ $(1 \leq k \leq K)$ capture the class-wise data variances of the training data. Therefore, we may make use of them to improve the classification accuracy. To be more specific, for the k-th class, we can first project the training bags as $\hat{X}_i = \mathbf{w}_k^T X_i$ $(1 \leq i \leq N)$ and the query bag as $\hat{X} = \mathbf{w}_k^T X$, then compute the corresponding ISD. It can be verified that the resulted ISD is

$$D(C_k, X_{i'}) = \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ikj}\ \hat{d}_k(\mathbf{x}_{ij}, X_{i'}) \quad , \tag{13}$$

where

$$\hat{d}_k(\mathbf{x}_{ij}, X_{i'}) = \left\|\mathbf{w}_k^T (\mathbf{x}_{ij} - \tilde{\mathbf{x}}_{ij})\right\|^2, \quad \forall\, i \in \pi_k, \tag{14}$$

and $\tilde{\mathbf{x}}_{ij}$ is the nearest neighbor of $\mathbf{x}_{ij}$ in $X_{i'}$. Thus, the ISD computed by Eq. (13) can be seen as C2B distance enhanced by both instance-specific and class-specific adaptations.

Label Prediction Using ISD

Given a test object, using the learned s_{ikj} by Algorithm 1, we can compute the ISDs $D(C_k, X_i)$ $(1 \leq k \leq K)$ from all the classes to the test bag using either Eq. (4) or Eq. (13). Sorting $D(C_k, X)$, we can easily assign labels to the test bag.

For a single-label multi-instance data set, in which each object belongs to one and only one class, we can assign X to the class with minimum ISD:

$$l(X) = \arg\min_k D(C_k, X) \quad . \tag{15}$$

For a multi-label multi-instance data set, in which an object can be associated with more than one class label, we need a threshold to make prediction (Wang, Huang, and Ding 2009; Wang, Ding, and Huang 2010a). For every class, we learn a threshold from the training data as following:

$$b_k = \sum_{i=1}^{N} Y_{ik} D(C_k, X_i) \,/\, \sum_{i=1}^{N} Y_{ik}, \tag{16}$$

which is the average ISD from the k-th class to all its training bags. Thus we can classify X by the following rule: assign X to the k-th class if $D(C_k, X) < b_k$, and not otherwise.

Experimental Results

In this section, we empirically evaluate the proposed method on five benchmark multi-instance data sets and two real world applications. In all our evaluations, we set $p = 0.01$.

Benchmark Tasks

We first evaluate the proposed ISD method on five benchmark data sets popularly used in studies of MIL, including **Musk1**, **Musk2**, **Elephant**, **Fox** and **Tiger**. All these data sets have only two classes, therefore binary classification is conducted on each of them respectively. Musk1 contains 47 positive and 45 negative bags, Musk2 contains 39 positive and 63 negative bags, each of the other three data sets contains 100 positive and 100 negative bags. More details of the data sets can be found in (Dietterich, Lathrop, and Lozano-Pérez 1997; Andrews, Hofmann, and Tsochantaridis 2002).

We evaluate the proposed ISD method, which uses Eq. (4) and solves J_2 in Eq. (6), via standard 5-fold cross-validation, where the parameter γ in Eq. (6) is fine tuned by an internal 5-fold cross-validation on the training data of each of the 5 trials in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 1, \ldots, 10^4, 10^5\right\}$. In addition, we also evaluate the following variations of the proposed method: (A) ISD without SCs by using Eq. (3), denoted as "C2B", in which no learning is involved; (B) ISD learned by using least square objective in Eq. (5), denoted as "ISD-LS"; (C) ISD with class-specific adaptations using

Table 1: Classification accuracy on benchmark data sets.

Method	Musk1	Musk2	Elephant	Fox	Tiger
C2B	0.828	0.767	0.739	0.604	0.807
ISD-LS	0.840	0.781	0.755	0.620	0.816
ISD-CD	0.851	0.781	**0.779**	**0.635**	0.847
ISD	**0.853**	**0.790**	0.768	0.632	**0.853**
DD	0.785	0.732	0.718	0.563	0.776
EM-DD	0.769	0.740	0.723	0.583	0.791
MIMLSVM	0.810	0.769	0.747	0.601	0.803
miGraph	0.833	0.770	0.751	0.606	0.810
MIMLSVM+	0.831	0.772	0.750	0.610	0.808

Eq. (13), denoted as "ISD-CD". The average classification accuracies are reported in Table 1, which also shows the performances of several most recent MIL methods including (1) Diversity Density (DD) method (Maron and Ratan 1998), (2) EM-DD method (Zhang and Goldman 2002), (3) MIMLSVM method (Zhou and Zhang 2007), (4) miGraph method (Zhou, Sun, and Li 2009) and (5) MIMLSVM+ method (Li et al. 2011). The parameters of these methods are set as their optimal according to the original papers.

The results in Table 1 show that our methods consistently outperform other compared methods, sometimes very significantly, which demonstrate their effectiveness in MIL. Besides, the C2B method is not as good as the other three variations of the proposed method. This is consistent with the theoretical formulations in that C2B method is a lazy learning method without involving any learning process and does not take into account the special properties of MIL. In addition, the proposed ISD method is better than ISD-LS method. The latter uses least square loss function thereby by nature is not able to handle the outliers caused by the weak label associations. Finally, ISD-CD exhibits the best performance in two test data sets, which confirms the usefulness of incorporating class-specific adaptations.

Image Categorization Task

Image categorization is one of the most successful applications of MIL thanks to the recent advances in image representation techniques using semi-local, or patch-based, features, such as SIFT and geometric blur. These algorithms choose a set of patches in an image, and for each patch compute a fixed-length feature vector. This gives a natural multi-instance representation of an image data set, *i.e.*, a set of vectors per image, where the size of the set can vary from image to image. Therefore we evaluate the proposed ISD method in the image categorization task.

We use **PASCAL VOC 2010** data set (Everingham et al. 2010) in our evaluations, which contains 13321 images with 20 classes. Each image contains several objects, each of which is denoted by a rectangle bounding box as shown in the training and query images in Figure 1. The region in each bounding box is considered as an instance, and a set of low-level features are extracted from the region including region size, color correlogram, color moments, wavelet texture and shape descriptor following (Chen and Wang 2004). In order to run the experiments on contemporary personal computers, we randomly select images from the data set, such that at least 100 images are selected for each class, which leads to 1864 images used in our experiments.

Because PASCAL VOC 2010 data set is a multi-label data set, we measure the classification performances of the compared methods using five widely used multi-label evaluation metrics, as shown in Table 2, where "↓" indicates "the small the better" while "↑" indicates "the bigger the better". Details of these evaluation metrics can be found in (Schapire and Singer 2000).

We evaluate the proposed ISD method and its variations by comparing them against the same MIL methods used in benchmark task evaluations. Because DD, DD-SVM and miGraph methods are single-label classification methods, we conduct binary classification on every class of the data set using the one-*vs*.-rest strategy. We still use standard 5-fold cross-validation to evaluate the compared methods and fine tune the parameters in a same way as before. The average classification performance (mean $\pm$ standard deviation) over the 5 trials of the experiments are reported in Table 2, which again clearly demonstrate the advantages of the proposed methods in multi-instance multi-label classification.

Finally, we study the SCs learned for the instances in the super-bags by our method. In Figure 1, we show the SCs for the two instances of the top left training image in PASCAL VOC 2010 data set. It shows that a same object may have different SCs when it is in different super-bags. To be more specific, the ship instance in region A of the image has comparably higher SC (0.493) than that (0.097) of the person instance in region B when considering "ship" class. In contrast, when it is in the super-bag of "person", its SC (0.018) is lower than that (0.532) of the person instance. These observations are consistent with our intuitions and theoretical analysis, because the ship instance contributes considerably large in characterizing the "ship" semantic concept, while it contributes much less, or even possibly harmful, in characterizing the "person" concept. The same observations can also be seen on almost all the training images, which are not shown due to space limit. These results provide a concrete evidence of proposed ISD method's capability in revealing the semantic insight of multi-instance data.

Text Categorization Task

We further evaluate the proposed methods in one more real world application of text categorization. We use the MIL text data set published in (Zhang and Zhou 2009), which is derived from Reuters-21578 collection. The seven most frequent categories are used, and 2000 bags are generated, in which each instance is represented by a 243-dimensional feature vector. Following the same way as before, we evaluate the compared methods via standard 5-fold cross-validation. The results in Table 3 show that our methods are more effective in text classification task.

Conclusions

In this paper, we proposed a novel Instance Specific Distance (ISD) method for Multi-Instance Learning (MIL). Taking into account the challenges of MIL including Bag-to-Bag (B2B) distance between multi-instance data objects and

Table 2: Comparison of multi-label classification performances (mean ± std) on PASCAL VOC 2010 image data set.

Method	Hamming loss ↓	One-error ↓	Coverage ↓	Rank loss ↓	Average precision ↑
C2B	0.185 ± 0.017	0.322 ± 0.027	0.998 ± 0.073	0.180 ± 0.012	0.481 ± 0.029
ISD-LS	0.153 ± 0.011	0.286 ± 0.021	0.973 ± 0.070	0.165 ± 0.014	0.491 ± 0.025
ISD-CD	0.131 ± 0.010	0.274 ± 0.019	$\mathbf{0.895 \pm 0.062}$	$\mathbf{0.156 \pm 0.011}$	0.511 ± 0.030
ISD	$\mathbf{0.129 \pm 0.012}$	$\mathbf{0.263 \pm 0.017}$	0.904 ± 0.067	0.158 ± 0.013	$\mathbf{0.521 \pm 0.031}$
DD	0.199 ± 0.023	0.387 ± 0.020	1.243 ± 0.075	0.202 ± 0.018	0.431 ± 0.025
DD-SVM	0.195 ± 0.021	0.367 ± 0.024	1.142 ± 0.076	0.191 ± 0.016	0.460 ± 0.020
MIMLSVM	0.180 ± 0.018	0.349 ± 0.029	1.064 ± 0.084	0.181 ± 0.014	0.479 ± 0.026
miGraph	0.169 ± 0.010	0.306 ± 0.019	1.015 ± 0.069	0.179 ± 0.012	0.480 ± 0.023
MIMLSVM+	0.175 ± 0.013	0.321 ± 0.025	0.997 ± 0.076	0.175 ± 0.010	0.484 ± 0.021

Table 3: Comparison of multi-label classification performances (mean ± std) on text (Reuters-21578 collection) data set.

Method	Hamming loss ↓	One-error ↓	Coverage ↓	Rank loss ↓	Average precision ↑
C2B	0.141 ± 0.007	0.306 ± 0.014	0.979 ± 0.081	0.170 ± 0.012	0.281 ± 0.016
ISD-LS	0.115 ± 0.004	0.255 ± 0.010	0.911 ± 0.073	0.154 ± 0.013	0.296 ± 0.020
ISD-CD	$\mathbf{0.103 \pm 0.003}$	$\mathbf{0.241 \pm 0.009}$	$\mathbf{0.869 \pm 0.061}$	0.147 ± 0.010	0.311 ± 0.019
ISD	0.106 ± 0.005	0.250 ± 0.013	0.870 ± 0.071	$\mathbf{0.145 \pm 0.011}$	$\mathbf{0.317 \pm 0.019}$
DD	0.160 ± 0.011	0.319 ± 0.017	1.098 ± 0.076	0.198 ± 0.020	0.261 ± 0.018
DD-SVM	0.151 ± 0.008	0.307 ± 0.018	0.972 ± 0.081	0.186 ± 0.017	0.270 ± 0.016
MIMLSVM	0.132 ± 0.006	0.294 ± 0.012	0.945 ± 0.074	0.167 ± 0.013	0.285 ± 0.017
miGraph	0.121 ± 0.004	0.276 ± 0.015	0.927 ± 0.069	0.159 ± 0.012	0.290 ± 0.021
MIMLSVM+	0.117 ± 0.004	0.268 ± 0.016	0.917 ± 0.067	0.153 ± 0.012	0.292 ± 0.016

the weak label association, we classified multi-instance data using ISD computed by Class-to-Bag (C2B) distance with further consideration on the relevances of training instances with respect to their labeled classes. In order to address the outliers inherent in multi-instance data, we learned ISD by solving an ℓ_{0+}-norm minimization problem. We presented an efficient algorithm to solve the problem, together with the rigorous proof of its convergence. We conducted extensive empirical studies on five benchmark multi-instance data sets and two real world multi-instance applications of image and text categorizations. Promising experimental results demonstrated the effectiveness of the our method.

Acknowledgments

This research was supported by NSF-CNS 0923494, NSF-IIS 1041637, NSF-CNS 1035913.

References

Andrews, S.; Hofmann, T.; and Tsochantaridis, I. 2002. Multiple instance learning with generalized support vector machines. In *NIPS*.

Boiman, O.; Shechtman, E.; and Irani, M. 2008. In defense of nearest-neighbor based image classification. In *CVPR*.

Chen, Y., and Wang, J. 2004. Image categorization by learning and reasoning with regions. *JMLR* 5:913–939.

Dietterich, T.; Lathrop, R.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89(1-2):31–71.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.

Li, Y.; Ji, S.; Kumar, S.; Ye, J.; and Zhou, Z. 2011. Drosophila Gene Expression Pattern Annotation through Multi-Instance Multi-Label Learning. *ACM/IEEE TCBB*.

Maron, O., and Ratan, A. 1998. Multiple-instance learning for natural scene classification. In *ICML*.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*.

Schapire, R., and Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization. *Machine learning* 39(2):135–168.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *Jouranl of Royal Statistics Society B.* 58:267–288.

Wang, H.; Ding, C.; and Huang, H. 2010a. Multi-Label Classification: Inconsistency and Class Balanced K-Nearest Neighbor. In *AAAI*.

Wang, H.; Ding, C.; and Huang, H. 2010b. Multi-label linear discriminant analysis. In *ECCV*.

Wang, Z.; Hu, Y.; and Chia, L. 2010. Image-to-Class Distance Metric Learning for Image Classification. In *ECCV*.

Wang, H.; Huang, H.; and Ding, C. 2009. Image annotation using multi-label correlated Green's function. In *ICCV*.

Xiang, S.; Nie, F.; and Zhang, C. 2008. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition* 41(12):3600–3612.

Zhang, Q., and Goldman, S. 2002. EM-DD: An improved multiple-instance learning technique. In *NIPS*.

Zhang, M., and Zhou, Z. 2009. M^{3}MIML: A maximum margin method for multi-instance multi-label learning. In *ICDM*.

Zhou, Z., and Zhang, M. 2007. Multi-instance multi-label learning with application to scene classification. In *NIPS*.

Zhou, Z.; Sun, Y.; and Li, Y. 2009. Multi-instance learning by treating instances as non-I.I.D. samples. In *ICML*.