MILEAGE: Multiple Instance LEArning with Global Embedding

Dan Zhang DANZHANG@FB.COM
Facebook Incorporation, Menlo Park, CA 94025

Jingrui He JINGRUI.HE@GMAIL.COM
Computer Science Department, Stevens Institute of Technology, Hoboken, NJ 07030

Luo Si LSI@CS.PURDUE.EDU
Computer Science Department, Purdue University, West Lafayette, IN 47907

Richard D. Lawrence RICKLAWR@US.IBM.COM
IBM T.J. Watson Research Center, Yorktown Heights, NY 10562

Abstract

Multiple Instance Learning (MIL) generally represents each example as a collection of instances such that the features for local objects can be better captured, whereas traditional learning methods typically extract a global feature vector for each example as an integral part. However, there is limited research work on investigating which of the two learning scenarios performs better. This paper proposes a novel framework – *Multiple Instance LEArning with Global Embedding (MILEAGE)*, in which the global feature vectors for traditional learning methods are integrated into the MIL setting. MILEAGE can leverage the benefits derived from both learning settings. Within the proposed framework, a large margin method is formulated to adaptively tune the weights on the two different kinds of feature representations (i.e., global and multiple instance) for each example and trains the classifier simultaneously. Some important properties of the proposed method are analyzed. Experiments on image, text and a novel application – Insider Threat Detection are conducted to demonstrate the advantages of the proposed method.

1. Introduction

Traditional learning methods usually consider each example as one non-separable entity, and represent the whole content of the example by one feature vector. However, the semantic meanings of an example sometimes vary among its constituent parts. *Multiple Instance Learning (MIL)* has been proposed to deal with problems whose output information is only known for bags of items/instances, as opposed to for each example. More precisely, in a MIL setting, each example/bag is divided into several different parts/instances. The labels are assigned to bags, rather than individual instances. A bag is labeled as positive if it contains more than one positive instance; otherwise it is labeled as negative. In this paper, for each example, the feature vector extracted by using the same way as we do for traditional non-MIL methods (i.e., treating each example as an integral entity) is referred to as the ***global representation*** of this example, while its ***local representation*** is a set of instances extracted for each part of this example, as in MIL. To some extent, the global representation for each example can also be considered as its bag level features[1]. Numerous methods have been developed for MIL classification (Andrews et al., 2003; Dietterich et al., 1998; Rahmani & Goldman, 2006; Kim & la Torre, 2010; Zhang et al., 2011) and its variants, such as outlier detection (Wu et al., 2010), online learning (Babenko et al., 2011), ranking (Hu et al., 2008), etc. These methods have been widely employed in areas such as text mining (Andrews et al., 2003) and localized content based image retrieval (LCBIR) (Rahmani & Goldman, 2006).

Most previous MIL methods focused on improving classification performance under local representation. However, few of them investigated whether the local representation is always better than the global one. This problem has posed a big challenge for researchers to decide what kind of algorithms should be used when facing real world applications.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[1] In many cases, the global representation can be roughly considered as a convex combination of local ones.

In (Ray & Craven, 2005), the authors compared the performances of traditional and MIL methods. However, their work is still based on the local representation, and adapts the traditional learning methods to the local representation.

Although rarely studied, it is intuitive that the true positive rates in positive bags could affect the performances of local and global representations significantly. This is because if the true positive rate in a positive bag is low, then its global representation will be dominated by the irrelevant parts of this example, while methods based on local representation could pick the true positive instances for training. On the contrary, if an example has few irrelevant parts, then the global representation tends to be more informative than the local one, since methods based on local representations normally focus on some local parts of each example. This intuition can also be verified empirically by the experiments conducted in Section 4.1. When incorporating this intuition into real applications, the major challenge is how to learn for each training example, whether local representation is better or global one tends to prevail.

To solve this challenge, a novel research framework – Multiple Instance LEArning with Global Embedding (MILEAGE) is proposed. MILEAGE leverages the benefits from both local and global representations such that in general it can achieve a better performance than both MIL and traditional learning methods. From another perspective, local and global feature representations can be treated as two information sources, and each of them carries some auxiliary information to improve classification performance, which is similar to the basic motivation of multi-view learning methods (Joachims et al., 2001). To solve the proposed framework, a novel method is designed by adaptively tuning the importance of the two different representations. It is based on the intuition that local representation tends to perform better when the positive ratio is small. An iterative method is employed to solve the derived optimization problem. To accelerate the optimization speed, inspired by (Fuduli et al., 2004), we adapt the bundle method to solve the resulting non-convex non-smooth problem by explicitly considering the convex regularization and the non-convex loss terms. Some discussions and theoretical analysis have been provided on important properties such as convergence rate and generalized error rate of the proposed method. Experiments on image, text datasets and a novel application – Insider Threat Detection, demonstrate the advantages of the proposed method.

2. Methodology

2.1. Problem Statement and Notation

Suppose a set of examples: $\mathcal{D} = \{(\mathbf{B}_i, \mathbf{B}_{i*}, Y_i), i = 1, \ldots, n\}$ are given, where $\mathbf{B}_i \in \mathcal{R}^{d\times 1}$ denotes the global representation for the i-th example and $Y_i \in \{1, -1\}$ is its binary label. Along with the global feature representation, for each example, its local feature representations, i.e., instances for different parts of this example, are also available (The notions of global and local representations are defined in Section 1). The instances in the i-th bag are denoted as: $\mathbf{B}_{i*} = \{\mathbf{B}_{i1}, \mathbf{B}_{i2}, \ldots, \mathbf{B}_{in_i}\} \in \mathcal{R}^{d\times n_i}$ [2], and n_i is the number of instances in the i-th bag. Throughout the paper, subscript $*$ means $j = 1, \ldots, n_i$. Given an unlabeled example $\mathbf{B}_u$ and its associated local representations, i.e., $\mathbf{B}_{u*}$, the objective of Multiple Instance LEArning with Global Embedding (MILEAGE) is to design a function $f : (\mathbf{B}_u, \mathbf{B}_{u*}) \to \mathcal{R}$, such that the classification on this unlabeled example is accurate. If $f(\mathbf{B}_u, \mathbf{B}_{u*}) > 0$, this example is classified as positive and otherwise negative.

2.2. Method

This section proposes a large margin formulation of the MILEAGE framework. Some important properties of the formulation will be presented in Section 3.

For each bag, a weight variable is introduced to balance the importance of the two representations. The weight is decided by both the prior knowledge from the positive ratio for each bag and the fitness of the data. Without loss of generality, given a specific example $\mathbf{B}_i$ and its associated instances $\mathbf{B}_{i*}$, the classifier takes the following form:

$$f(\mathbf{B}_i, \mathbf{B}_{i*}) = \lambda_i \max_j \mathbf{w}^T \mathbf{B}_{ij} + (1 - \lambda_i)\mathbf{w}^T \mathbf{B}_i, \qquad (1)$$

where $1 \geq \lambda_i \geq 0$ is the convex combination coefficient for the i-th example, $\mathbf{w} \in \mathcal{R}^{d\times 1}$ is the linear classifier and we assume that the bias has already been absorbed into feature vectors. $\max_j \mathbf{w}^T \mathbf{B}_{ij}$ is the output from the local representation of the i-th example[3], whereas $\mathbf{w}^T \mathbf{B}_i$ is the output from its global representation. $f(\mathbf{B}_i, \mathbf{B}_{i*})$ balances these two outputs through the weight λ_i. From a Bayesian perspective, given a dataset, the logarithm of the posterior distribution for $\mathbf{w}$ and λ can be written as follows:

$$\log P(\mathbf{w}, \lambda|\mathcal{D}) \propto \log P(\mathcal{D}|\mathbf{w}, \lambda)P(\mathbf{w}) \prod_{i=1}^{n} P(\lambda_i), \qquad (2)$$

where $\lambda = [\lambda_1, \ldots, \lambda_n]$. Here, we assume that the examples are i.i.d. generated[4]. $P(\mathbf{w})$ follows the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. $P(\lambda_i)$ follows the Beta distribution with $\mathrm{beta}(\gamma e^{-\mu r_i}, \gamma e^{-\mu(1-r_i)})$, where μ and γ are the hyperparameters and partially control the mean and skewness of the distribution. $r_i \in [0, 1]$ is the prior knowledge on the positive ratio for the i-th bag, and can be obtained

[2] We assume that the local and global representations share the same feature space. But the proposed formulation can be extended to the case when their feature spaces are different.

[3] The output of each example in MIL is normally decided by the instance that appears to be most positive under a classifier $\mathbf{w}$ (Andrews et al., 2003)

[4] The proposed formulation can also be extended to the cases where examples are not i.i.d. generated in a similar way as (Zhou et al., 2009)

through various ways. For example, r_i can be simply set to 0.5 if no prior knowledge is available. In practice, a preliminary classifier can be trained beforehand by using SVM on $\{(\mathbf{B}_i, Y_i), i = 1, \ldots, n\}$. Then, r_i can be estimated by applying this classifier on the instances in each bag. It is clear that $E(\lambda_i) = e^{-\mu r_i}/(e^{-\mu r_i} + e^{-\mu(1-r_i)})$. Given $\mathbf{w}$ and λ, the probability of generating a dataset $\mathcal{D}$ can be described by the hinge loss as: $P(\mathcal{D}|\mathbf{w}, \lambda) \propto \prod_{i=1}^{n} e^{-C \max\{0, 1-Y_i(\lambda_i \max_j \mathbf{w}^T \mathbf{B}_{ij} + (1-\lambda_i)\mathbf{w}^T \mathbf{B}_i)\}}$, where C is a parameter. Then, maximizing Eq.(2) is equivalent to solving the following problem:

$$\begin{aligned} \min_{\mathbf{w}, \lambda, \xi_i \geq 0} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \big((\gamma e^{-\mu r_i} - 1)\log \lambda_i \\ & + (\gamma e^{-\mu(1-r_i)} - 1)\log(1-\lambda_i)\big) \qquad (3) \\ s.t.\ & \forall i \in \{1, \ldots, n\}, \\ & Y_i(\lambda_i \max_j \mathbf{w}^T \mathbf{B}_{ij} + (1-\lambda_i)\mathbf{w}^T \mathbf{B}_i) \geq 1 - \xi_i. \end{aligned}$$

This formulation is a non-convex optimization problem and cannot be solved directly. An iterative method is employed to solve this problem. In particular, for the k-th iteration, given $\mathbf{w}^{(k-1)}$, $\lambda_1, \ldots, \lambda_n$ can be updated by:

$$\begin{aligned} \min_{\lambda} \ & C\sum_{i=1}^{n} \max\{0, 1 - Y_i(\lambda_i \mathbf{w}^{(k-1)T} \mathbf{B}_{ij_i^{(k-1)}} + (1-\lambda_i)\mathbf{w}^{(k-1)T} \mathbf{B}_i)\} \\ & - \sum_{i=1}^{n} ((\gamma e^{-\mu r_i} - 1)\log \lambda_i + (\gamma e^{-\mu(1-r_i)} - 1)\log(1-\lambda_i)) \qquad (4) \end{aligned}$$

where $j_i^{(k-1)} = \arg\max_j \mathbf{w}^{(k-1)T} \mathbf{B}_{ij}$. The convexity of this objective function cannot be determined, since the signs of $(\gamma e^{-\mu r_i} - 1)$ and $(\gamma e^{-\mu(1-r_i)} - 1)$ are not clear. But some methods, such as the adapted subgradient method, can still be used to find its optimal or local optimal solution efficiently. Given λ from the previous step, $\mathbf{w}^{(k)}$ can be optimized by:

$$\begin{aligned} \min_{\mathbf{w}} \ & \frac{1}{2}\|\mathbf{w}\|^2 + \\ & C\sum_{i=1}^{n} \max\{0, 1 - Y_i(\lambda_i \max_j \mathbf{w}^T \mathbf{B}_{ij} + (1-\lambda_i)\mathbf{w}^T \mathbf{B}_i)\} \qquad (5) \end{aligned}$$

It is still a non-convex non-smooth optimization problem. But the form is much less complicated than that of problem (3). It can be solved through various ways, such as constrained concave-convex procedure (CCCP) (Yuille & Rangarajan, 2003). However, the computational cost for solving this problem is non-trivial. In several recent works, the bundle method has shown its superior performance in both efficiency and effectiveness over state-of-the-art methods (Joachims, 2006; Joachims et al., 2009; Smola et al., 2007; Teo et al., 2010). However, one major drawback for this method is that it can only be employed to solve convex optimization problems. In (Fuduli et al., 2004; Hare & Sagastizábal, 2010; Noll, 2012), several heuristics are employed to handle this issue for the bundle method. In this paper, inspired by (Fuduli et al., 2004), we further adapt the bundle method to solve this proposed optimization problem in the next section. Based on these updating schemes, problem (4) and problem (5) will be conducted iteratively until convergence.

It is clear that the proposed formulation is inductive on the classifier but transductive on λ_i. So, if we only need to predict the unlabeled instances in the unlabeled set, then we can directly apply the learned classifier. If the prediction is made on the bag level, on an unlabeled example $(\mathbf{B}_u, \mathbf{B}_{u*}), j = 1, \ldots, n_u$. Its hidden variable λ_u can be estimated as: $\lambda_u^* = E(\lambda_u|\mathbf{B}_u, \mathbf{B}_{u*}) = e^{-\mu r_u}/(e^{-\mu r_u} + e^{-\mu(1-r_u)})$, where r_u is the positive instance ratio within this bag estimated from the learned classifier $\mathbf{w}$. Then, $f(\mathbf{B}_u, \mathbf{B}_{u*}) = \lambda_u^* \max_j \mathbf{w}^T \mathbf{B}_{uj} + (1 - \lambda_u^*)\mathbf{w}^T \mathbf{B}_u$. If $f(\mathbf{B}_u, \mathbf{B}_{u*}) > 0$, the example is labeled as positive and otherwise it is labeled as negative.

2.3. Bundle Method for Non-Convex Non-Smooth Optimization

The traditional bundle method tries to find a set of cutting planes that could serve as the support/lower bounds of the original convex objective function. For non-convex optimization problems, however, these cutting planes could no longer serve as lower bounds of the objective functions, as shown in Fig.1. Some research works consider shifting of affine pieces downwards (Noll, 2012; Schramm & Zowe, 1992). However, the amount of the shifting appears arbitrary (Fuduli et al., 2004).

In this section, the bundle method, which is based on first order approximation, is adapted to solve problem (5). In particular, the intended objective function can be casted as the following framework:

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) = \Omega(\mathbf{w}) + R_{emp}(\mathbf{w}), \qquad (6)$$

where $\Omega(\mathbf{w})$ is a non-negative convex differentiable regularizer, and $R_{emp}(\mathbf{w})$ is a non-convex non-smooth loss function. In problem (5), $\Omega(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ and $R_{emp}(\mathbf{w}) = C\max\{0, 1 - Y_i(\lambda_i \max_j \mathbf{w}^T \mathbf{B}_{ij} + (1-\lambda_i)\mathbf{w}^T \mathbf{B}_i)\}$.

Our proposed method handles this non-convex non-smooth problem in an iterative way and exhibits a kind of both convex and nonconvex behavior relative to the current point in the iterative procedure. More precisely, for the t-th iteration of the bundle method, it maintains two different sets of cutting planes, i.e., $I_+ \triangleq \{j|\alpha_j^{(t)} \geq 0\}$, $I_- \triangleq \{j|\alpha_j^{(t)} < 0\}$, where $j = 1, \ldots, t-1$

$$\alpha_j^{(t)} \triangleq R_{emp}(\mathbf{w}^{(t-1)}) - R_{emp}(\mathbf{w}^{(j)}) - \mathbf{g}_j^T(\mathbf{w}^{(t-1)} - \mathbf{w}^{(j)}) \qquad (7)$$

Here, $\mathbf{g}_j \in \partial_{\mathbf{w}} R_{emp}(\mathbf{w}^{(j)})$[5]. Then, the following two sets

[5]For simplification, we abused the superscript. Please note that in this section, the superscript t denotes the t-th iteration in the bundle method.

of affine functions are defined as:

$$\Delta^+(\mathbf{w}) \triangleq \max_{j\in I_+} \mathbf{g}_j^T(\mathbf{w}-\mathbf{w}^{(t-1)}) - \alpha_j^{(t)},$$
$$\Delta^-(\mathbf{w}) \triangleq \min_{j\in I_-} \mathbf{g}_j^T(\mathbf{w}-\mathbf{w}^{(t-1)}) - \alpha_j^{(t)}. \quad (8)$$

It is clear that $\Delta^+(\mathbf{w})$ is an approximation of $R_{emp}(\mathbf{w}) - R_{emp}(\mathbf{w}^{(t-1)})$, while $\Delta^-(\mathbf{w})$ is its locally pessimistic estimation. These approximations are only locally valid around the local minimal point. Here, the meanings of $\alpha_j^{(t)}$ and the locality property can be shown in Fig.1. Therefore, during each iteration, the new optimal point should tradeoff minimizing $\Delta^+(\mathbf{w})$ and proximity $\|\mathbf{w}-\mathbf{w}^{(t-1)}\|$ with the constraint $\Delta^+(\mathbf{w}) \le \Delta^-(\mathbf{w})$ as follows:

$$\min_{\mathbf{w},\zeta} \quad P(\mathbf{w},\gamma^{(t)}) = \gamma^{(t)}(\zeta+\Omega(\mathbf{w})) + \frac{1}{2}\|\mathbf{w}-\mathbf{w}^{(t-1)}\|^2 \quad (9)$$
$$s.t. \quad \zeta \ge \mathbf{g}_j^T(\mathbf{w}-\mathbf{w}^{(t-1)}) - \alpha_j^{(t)}, \ \ j\in I_+,$$
$$\zeta \le \mathbf{g}_j^T(\mathbf{w}-\mathbf{w}^{(t-1)}) - \alpha_j^{(t)}, \ \ j\in I_-,$$

where $\gamma^{(t)}$ is the non-negative proximity control parameter for the t-th iteration that balances the objective function value and the proximity of the updated point. This problem can be solved efficiently through its dual form, since both of the sets I_+ and I_- are small. Suppose $\mathbf{w}^{(t)} = \arg\min_{\mathbf{w}} P(\mathbf{w},\gamma^{(t)})$. If not computationally expensive, a line search can be performed between $\mathbf{w}^{(t)}$ and $\mathbf{w}^{(t-1)}$ on $F(\mathbf{w})$ such that a better solution can be found.

If the optimal solution can result in a drastic decrease in the objective function $F(\mathbf{w})$, it is called a serious step and the optimal solution for $\mathbf{w}$ will be updated. Otherwise, it is considered as a null step, the optimal solution for the previous step is kept, and the proximity parameter will shrink for a better solution. If $\|\mathbf{w}^{(t)}-\mathbf{w}^{(t-1)}\|$ is less than a predefined threshold θ, the proximity parameter will also shrink to do a more thorough search within that region.

The classic bundle method usually checks whether the difference between the objective function value and the cutting plane function value is less than a threshold. If so, the iteration terminates. Here, this strategy cannot be used because the cutting planes of the non-convex function cannot be considered as the lower bounds for the original objective function any more. In the proposed method, during each iteration, two stopping criteria will be checked. The first stopping criteria is to check whether $\gamma^{(t)}$ is smaller than a specified threshold ϵ_1. This is because although we hope that the new updated point should fall within a small region of $\mathbf{w}^{(t-1)}$, if $\gamma^{(t)}$ becomes too small, $\mathbf{w}^{(t)}$ is unlikely to deviate too much from $\mathbf{w}^{(t-1)}$, and the results will not be meaningful. An extreme example is if $\gamma^{(t)} = 0$, then $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)}$. The second stopping criteria is to check whether $0 \in \partial F(\mathbf{w}^{(t)})$, i.e., whether $\mathbf{w}^{(t)}$ can be considered as a stationary point for $F(\mathbf{w})$. In practice, we check whether $\|\mathbf{o}^*\|/F(\mathbf{w}^{(t)}) \le \delta$, where $\mathbf{o}^* = \min_{\mathbf{o}\in conv\{\mathbf{g}_j|j\in J_+\}} \|\mathbf{o}+\partial\Omega(\mathbf{w}^{(t)})\|$ and $J_+ = \{i \in I_+|\alpha_i^{(t)} \le \epsilon_2\}$. In particular,

$$\mathbf{o}^* = \mathbf{G}\upsilon^* + \partial\Omega(\mathbf{w}^{(t)}), \quad (10)$$

where $\mathbf{G}$ is a matrix with its columns being the subgradients $\mathbf{g}_j$ from J_+ and υ^* can be optimized by solving $\upsilon^* = \arg\min \upsilon^T\mathbf{G}^T\mathbf{G}\upsilon + 2(\partial\Omega(\mathbf{w}^{(t)}))^T\mathbf{G}\upsilon \ \ s.t. \ \upsilon^T\mathbf{1} = 1, \ \ \upsilon \ge 0$.

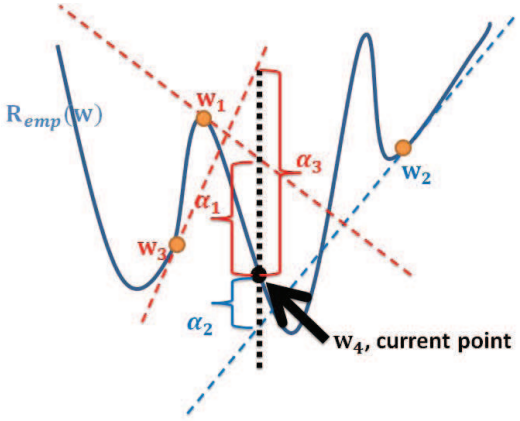


Figure 1. Approximation of $R_{emp}(\mathbf{w})$ at $\mathbf{w}_4$. The cutting planes from other points either over or underestimate the value at and in the vicinity of $\mathbf{w}_4$, and the sign of α_i will not change in the vicinity of $\mathbf{w}_4$ ($\alpha_1 < 0$, $\alpha_2 > 0$, $\alpha_3 < 0$). Based on this locality characteristic, we adapted the bundle method in Section 2.3.

3. Discussions and Theoretical Analysis

The proposed bundle method is summarized in Table 1. It is clear that the major advantage of the proposed method over (Fuduli et al., 2004) is that the proposed method better exploits the structure of the objective function by treating the convex and non-convex parts separately. It therefore eliminates the unnecessary first order approximation for the convex part. In this way, theoretically the cutting plane approximation for the whole objective function is more accurate than the one used in (Fuduli et al., 2004).

In (Bergeron et al., 2012), the authors directly applied (Fuduli et al., 2004) to MIL. However, there are several major differences between these two papers. 1. (Bergeron et al., 2012) only focuses on the traditional MIL, and can not be used to solve MILEAGE. 2. By directly employing (Fuduli et al., 2004), (Bergeron et al., 2012) does not treat the convex and non-convex parts separately either and therefore its first order approximation is less accurate than the one used in this paper.

In (Do & Artières, 2009), the non-convex formulation for hidden markov models is also solved by adapting the bundle method to the non-convex case, and treating the convex and non-convex parts separately. The adapted method

Input: 1. The objective function: $\Omega(\mathbf{w}) + R_{emp}(\mathbf{w})$. 2. Parameters: descent coefficient: $m = 0.1$, initial proximity control parameter $\gamma^{(1)} = 1$, deviation parameters $\epsilon_1 = 0.01$, $\epsilon_2 = 0.1$ and $\theta = 0.01$, decay coefficient $\eta = 0.9$, gradient precision $\delta = 0.01$. **Output:** $\mathbf{w}$.

1. Initialize $\mathbf{w}^{(1)}$, t=1
repeat:
2. $t = t + 1$.
3. Get $(\mathbf{w}^{(t)}, \zeta^{(t)})$ by solving the dual of problem (9).
4. **If** $F(\mathbf{w}^{(t)}) \leq F(\mathbf{w}^{(t-1)}) + m(\zeta^{(t)} + \Omega(\mathbf{w}^{(t)}) - \Omega(\mathbf{w}^{(t-1)}))$ **and** $\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\| \geq \theta$
5. $\gamma^{(t)} = \gamma^{(t-1)}$
6. **else**
7. $\gamma^{(t)} = \eta\gamma^{(t-1)}$
8. **If** $\gamma^{(t)} \leq \epsilon_1$, **then** exit.
9. **If** $F(\mathbf{w}^{(t)}) > F(\mathbf{w}^{(t-1)}) + m(\zeta^{(t)} + \Omega(\mathbf{w}^{(t)}) - \Omega(\mathbf{w}^{(t-1)}))$, **then**, $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)}$.
10. **end if**
11. $I_+ = \phi$, $I_- = \phi$
12. **for** j=1 . . . t
13. Evaluate $\alpha_j^{(t)}$ according to Eq.(7), if $\alpha_j^{(t)} \geq 0$, then, $I_+ = I + \bigcup j$; if $\alpha_j < 0$, then, $I_- = I_- \bigcup j$;
14. **end for**
15. Compute o^* according to Eq.(10). If $\|o^*\|/F(\mathbf{w}^{(t)}) \leq \delta$, then **exit**.
until algorithm terminates
16. $\mathbf{w} = \mathbf{w}^{(t)}$.

Table 1. The proposed bundle method for non-convex non-smooth optimization

is reasonable by tuning the cutting plane at each iteration according to the comparison with the previous "optimal" cutting plane. However, even with this tuning, the obtained cutting plane is still not able to serve as the lower bound of the objective function. On the contrary, the proposed method does not focus on looking for the lower bound, but some important local properties around each point.

Furthermore, based on the proposed bundle method, some important properties that explicitly consider both the convex and the non-convex parts of the objective function are analyzed in Theorem 1 and Theorem 2.

Theorem 1: Suppose $D = \max_t \Omega(\mathbf{w}^{(t)})$ and $R = \max_j \|\mathbf{g}_j\|$, then $-\frac{\gamma_0^2}{2}R^2 \leq P(\mathbf{w}^{(t)}, \gamma^{(t)}) \leq \gamma_0 D$. In solving problem (5), $-\frac{\gamma_0^2 C^2}{2}\max\{\max_{i,j}\|\mathbf{B}_{ij}\|^2, \max_i \|\mathbf{B}_i\|^2\} \leq P(\mathbf{w}^{(t)}, \gamma^{(t)}) \leq \gamma_0 D$.

Proof: Please refer to Supplemental Materials. □

Theorem 2: The bundle method terminates after at most $\log\frac{\epsilon_1}{\gamma_0}/\log(\eta) + \frac{2E\gamma_0}{m\theta^2}$ steps, given $\min R_{emp}(\mathbf{w}) + \Omega(\mathbf{w})$ is upper bounded by E. In solving problem (5), the algorithm terminates after at most $\log\frac{\epsilon_1}{\gamma_0}/\log(\eta) + \frac{2nC\gamma_0}{m\theta^2}$ steps.

Proof: Please refer to Supplemental Materials. □

Suppose the class of classifier satisfies $\|\mathbf{w}\| \leq B$ and λ are obtained from iterative updates. Since the proposed method can be easily extended to the kernel case, $\mathcal{F}_B$ is defined as: $\{f | f : (\mathbf{B}_i, \mathbf{B}_{i*}) \to \lambda_i \max_j \mathbf{w}^T\phi(\mathbf{B}_{ij}) + (1 - \lambda_i)\mathbf{w}^T\phi(\mathbf{B}_i), \|\mathbf{w}\| \leq B\}$, where ϕ is a nonlinear map with kernel function $K(\cdot,\cdot)$. The generalized error bound can be derived by the following theorems:

Theorem 3: The empirical Rademacher complexity of the functional space $\mathcal{F}_B$ on $\mathcal{D} = \{(\mathbf{B}_i, \mathbf{B}_{i*}, Y_i), i = 1, \ldots, n\}$ is upper bounded by: $\frac{2B}{n}\max_{\varphi_{ij}\geq 0, \varphi_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^n\sum_{j=1}^{n_i}\lambda_i^2\varphi_{ij}^2 K(\mathbf{B}_{ij}, \mathbf{B}_{ij})} + \frac{2B}{n}\sqrt{\sum_{i=1}^n(1-\lambda_i)^2 K(\mathbf{B}_i, \mathbf{B}_i)}$.

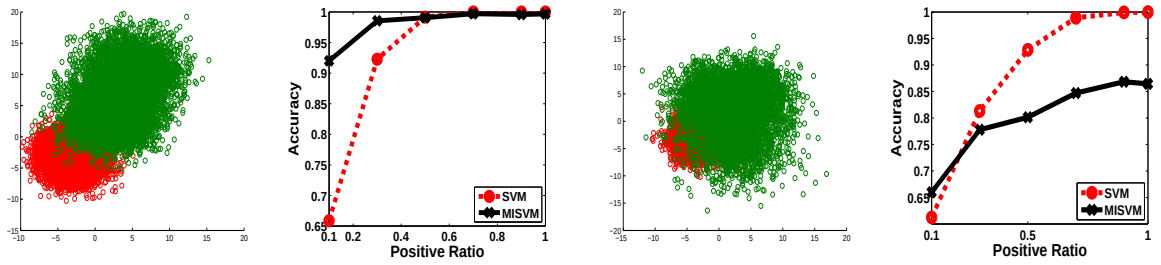
Proof: Please refer to Supplemental Materials. □

Theorem 4: Fix $\kappa \in (0, 1)$. Then, with probability at least $1 - \kappa$, every $f \in \mathcal{F}_B$ satisfies: $P(y \neq sign(f(\mathbf{B}_i, \mathbf{B}_{i*}))) \leq \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i(\lambda_i \max_j \mathbf{w}^T\mathbf{B}_{ij} + (1 - \lambda_i)\mathbf{w}^T\mathbf{B}_i)\} + \frac{2B}{n}\max_{\varphi_{ij}\geq 0, \varphi_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^n\sum_{j=1}^{n_i}\lambda_i^2\varphi_{ij}^2 K(\mathbf{B}_{ij}, \mathbf{B}_{ij})} + \frac{2B}{n}\sqrt{\sum_{i=1}^n(1-\lambda_i)^2 K(\mathbf{B}_i, \mathbf{B}_i)} + 3\sqrt{\frac{ln(2/\kappa)}{2n}}$.

Proof: This result can be got by applying Theorem 3 to Theorem 4.9 in (Shawe-Taylor & Cristianini, 2004). □

From Theorem 3 and Theorem 4, it can be seen that the derived Rademacher complexity and generalized error bound are related to both the local and global feature representations. Theorem 5 states the case when the Rademacher Complexity can be improved, compared with both local and global feature representations.

Theorem 5: Suppose $a \leq \lambda_i \leq \max\{\frac{C_2}{C_1}a, 1 - \frac{C_2}{C_1}(1 - a)\}, i = 1, \ldots, n$, $a \in [0, 1]$, where $C_1 = \frac{2B}{n}\max_{\varphi_{ij}\geq 0, \varphi_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^n\sum_{j=1}^{n_i}\varphi_{ij}^2 K(\mathbf{B}_{ij}, \mathbf{B}_{ij})}$



(a) (b) (c) (d)

Figure 2. Experiments with different positive ratios. The true positive (marked by red) and negative (marked by blue) instance distributions are shown in (a) and (c) respectively, with different amount of overlap. The positive bags are generated by extracting specified ratios of positive instances (as indicted in x-axis of (b) and (d)) and negative instances from the two instance distributions, while negative bags are composed of negative instances. SVM and its MIL variant – MISVM (Andrews et al., 2003), are used for comparisons. Here, for SVM, experiments are conducted on the averaged features in each bag. In (b) and (d), the averaged accuracy of 20 independent runs under different positive ratios are reported for datasets generated from (a) and (c) respectively.

and $C_2 = \frac{2B}{n}\sqrt{\sum_{i=1}^{n} K(\mathbf{B}_i, \mathbf{B}_i)}$, then, $\frac{2B}{n}\max_{\varphi_{ij}\geq 0, \varphi_i^T \mathbf{1}=1} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n_i} \lambda_i^2 \varphi_{ij}^2 K(\mathbf{B}_{ij}, \mathbf{B}_{ij})} + \frac{2B}{n}\sqrt{\sum_{i=1}^{n}(1-\lambda_i)^2 K(\mathbf{B}_i, \mathbf{B}_i)} \leq \max\{C_1, C_2\}$.

Proof: Please refer to Supplemental Materials. □

In Theorem 5, C_1 indicates the Rademacher Complexity derived from the local representation, while C_2 represents the Rademacher Complexity for the global representation. It can be concluded that, under some restrictions, the Rademacher Complexity of the proposed method is guaranteed to be less than the maximum one of the Rademacher Complexities for local and global representations.

4. Experiments

4.1. Synthetic Experiments

The synthetic dataset is designed to verify the intuitions conveyed in this paper, i.e., local representation works better when the true positive ratios in positive bags are lower, while global representation works better when the ratios are higher. In particular, we design two sets of experiments as shown in Fig.2. For each set of experiments, positive instances are generated from a Gaussian distribution, and negative ones are generated from another three Gaussian distributions, with different amounts of overlap, as shown in Fig.2(a) and Fig.2(c). Based on these two data distributions, for each set of experiments, 6 toy datasets are created with each positive bag containing a certain ratio of positive and negative instances and each negative bag containing all negative instances. Each bag contains 10 instances. For each dataset 1000 positive bags and 1000 negative ones are i.i.d. generated. SVM and its MIL variant – MISVM (Andrews et al., 2003) (We report the comparison results of these two methods, because their objective functions are the same except for the local and global representation part) are used for comparison, where the average feature representation of each bag is used as its global feature representation and used by SVM. For each experiment, 50% examples are randomly picked for training, and the rest for testing. The averaged results of 20 independent runs under different ratios of positive instances in positive bags are reported in Fig.2(b) and Fig.2(d) for datasets generated from Fig.2(a) and Fig.2(c) respectively, with the parameters tuned by 5-fold cross validation.

It can be partially concluded from the experiments that: (1) The local representation is not always better than the global one; (2) The local representation tends to perform better than the global one when the positive ratio is low; (3) There is no universally "good" positive ratio below which the local representation is definitely better than the global one. (4) It seems that if the amount of overlap between positive and negative distributions is high, the local representation is likely to be worse than that of the global representation[6].

4.2. Real Applications

These experiments are conducted on three datasets, i.e., an image dataset from Corel (Andrews et al., 2003), a text dataset from Reuters21578 [7] as well as newly proposed application – Insider Threat Detection. In MIL, MUSK (Dietterich et al., 1998) is also a commonly used benchmark dataset. But its performance is not reported here, because the meaning of the global representation for MUSK is not clear. In MUSK, instances represent different conformations of molecule. For each molecule, a set of con-

[6] Please note that although the overlap in Fig.2(c) looks heavy on the instance level, SVM can still attain almost 100% accuracy. This is because their global features are much better separated, since the average of the instance features greatly reduce the variance of the distributions.

[7] http://daviddlewis.com/resources/textcollections/reuters21578/

formations do not convey physical meanings in global representation. But for images and documents, each image or document itself can be considered a concrete object.

The Corel dataset is divided into three sub-datasets, i.e., Fox, Elephant and Tiger. For a detailed description of these three datasets, please refer to (Andrews et al., 2003). For each picture/example in Corel, the global feature vector is the average of the instances on all dimensions.

For Reuters21578, documents from 4 sub-categories, as well as some negative documents, are randomly picked. For each of the sub-dataset, after removing the stop words and stemming, tf-idf (Manning et al., 2008) features are extracted and processed by PCA (Berry & Castellanos, 2007). The resulting dimensionality is 249. For each document/bag, the global feature vector is extracted from the whole content; while the instance features are derived through a sliding window with fixed length (Andrews et al., 2003). For Reuters1, Reuters2, Reuters3, Reuters4, they contain $1602, 1256, 1100, 502$ bags, and $3006, 2181, 2249, 920$ instances, respectively.

For Insider Threat Detection (ITD), We obtained this real dataset from a big IT company. ITD is a project which is devoted to find the potential harmful insiders through analyzing their online behaviors, such as sending emails, login, logout, downloaded files, etc. In this dataset, some experts are hired to decide whether during each period (around 30 days), each person in the database did malicious things or not. Based on this, each online behavior is quantified as a feature value. However, it is highly possible that if a person did malicious things during a period, it does not mean that he did malicious things every day. Out of this motivation, the features for the online behaviors within one day is considered as an instance and the instances during each period is treated as a bag. If a person is known to do some malicious things in a specific period, then the corresponding collection of instances (days) is considered as a positive bag. Otherwise, this collection of instances will be considered as negative. The global feature representation for each bag is extracted from the corresponding period as a whole. The whole dataset contains 1000 negative bags and 166 positive bags, where each bag contains around 30 instances and each instance is represented by 32 features. On this dataset, due to the imbalance of the dataset, F1 score for the top 20 returned results is used here for measurement.

4.3. Comparison Results

In the proposed method, parameters C, γ and μ are set through 5-fold cross validation on the training set through the grids $2^{[-5:2:7]}$, $2^{[-4:2:8]}$ and $[0.1, 1, 10, 100]$ respectively. To show the advantages of the proposed large margin method, we compare it with several baseline methods, including traditional large margin methods, SVM-B, SVM-I, and multiple instance learning methods: Citation KNN (Wang & Zucker, 2000), MISVM (Andrews et al., 2003), miSVM (Andrews et al., 2003), MILES (Chen et al., 2006), and ISMIL (Fu & Robles-Kelly, 2009).

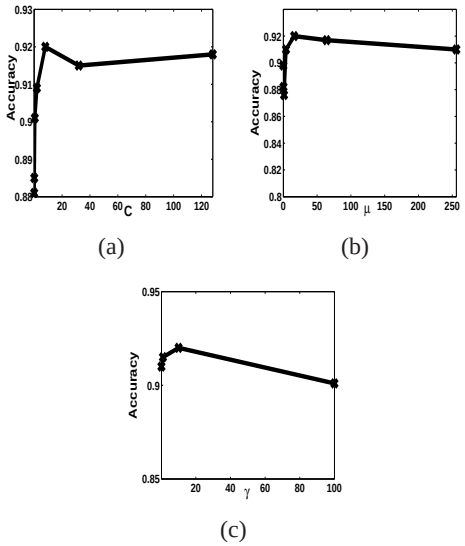


Figure 3. Parameter Sensitivity Comparisons

For SVM-B, SVM is used on the bag/global features for training and prediction. For SVM-I, the bag labels are assigned to their corresponding instances, and SVM is used on these labeled instances. For each unlabeled bag, if at least one of its instances is labeled as positive by SVM-I, then its bag label is positive. Otherwise, it is negative. As for MIL methods, Citation KNN is an adaptation of traditional K nearest neighbor to MIL. MISVM and miSVM are two large margin multiple instance classification methods, derived from SVM. MILES tries to represent each bag by using one feature vector, and then design a classifier based on that. For ISMIL, the algorithm maps bags into a space spanned by some selected instances, and designs a classifier based on that. The parameters of the baseline methods are also tuned by 5-fold cross validation. For the large margin methods, for the fair of comparison, only linear classifiers are used.

The average accuracy of 20 independent runs are reported in Table 2 and 3. For each experiment, 90% examples are randomly sampled as training examples, while the remaining ones are used for testing. It is clear that MIL methods are not better than the traditional learning methods on all of these datasets, which further verifies that the local representation for MIL may not be always better than the global representation. From these experimental results, in most cases, MILEAGE shows the best performance. This is because MILEAGE takes advantage of both local and global representations adaptively. These two different rep-

	Image			Text				Insider Threat Detection
	Fox	Elephant	Tiger	Reuters1	Reuters2	Reuters3	Reuters4	ITD
MILEAGE	**64.5**	**84.5**	**84.0**	90.3	**92.9**	**93.2**	**91.6**	**0.495**
SVM-B	53.8	83.0	76.0	90.5	92.0	91.1	88.2	0.397
SVM-I	56.5	71.2	72.5	88.5	92.4	88.9	89.4	0.401
CitationKNN	60.5	81.5	82.0	86.5	86.7	80.9	81.4	0.319
MISVM	59.0	78.3	81.7	**90.6**	91.9	91.3	90.4	0.474
miSVM	57.5	81.0	78.3	88.2	91.7	90.5	86.6	0.444
MILES	62.0	81.9	77.5	88.9	91.7	92.2	88.4	0.469
ISMIL	61.6	82.0	78.9	88.0	90.3	91.5	89.3	0.417

Table 2. Accuracy Comparisons (%) on Image and Text datasets, and F1 Comparisons on Insider Threat Detection. On ITD dataset, F1 score for the top 20 returned results is used here for measurement due to the imbalance of the dataset.

	Image			Text				Insider Threat Detection
	Fox	Elephant	Tiger	Reuters1	Reuters2	Reuters3	Reuters4	ITD
MILEAGE	26.2	48.7	58.3	254.8	225	74.5	53.9	12.6
SVM-B	0.06	0.01	0.02	0.7	0.7	0.4	0.2	0.8
SVM-I	0.8	0.5	0.4	0.9	1.2	0.8	0.7	0.7
CitationKNN	42.7	48.5	36.2	166.6	90.4	80.0	14.3	93.8
MISVM	27.9	13.7	13.5	242.3	431.2	309.3	165.5	4.6
miSVM	25.3	19.3	3.8	17.3	10.5	7.5	1.2	5.8
MILES	26.6	30.1	23.4	476.3	236.4	201.0	17.0	2308
ISMIL	9.5	11.3	10.0	210.2	100.3	62.1	10.8	58.4

Table 3. Time Comparisons (in seconds)

resentations can be considered as two different information sources, and both of them convey some useful information in improving the performance.

For time comparisons, the proposed method is comparable with most of the other MIL methods. The efficiency of the proposed bundle method plays an important role. For example, MISVM needs to solve a non-convex problem similar to problem (5) only once for each experiment, but the proposed method needs to solve problems (4) and (5) for around 15 times before convergence. So, the average amount of time needed for each independent execution of the bundle method is small, compared with that of MISVM. On the other side, traditional learning methods such as SVM-B and SVM-I tend to be more efficient because they can easily apply convex optimization methods such as Sequential Maximization Optimization to their convex objective functions once with only one kind of representations. But the proposed MILEAGE framework generate more accurate results in most cases due to the more realistic non-convex setting of both global representations and local representations.

To show the robustness of the proposed method, some parameter sensitivity experiments are conducted on C, μ, γ, and shown in Fig.3. The averaged experiments of 20 independent runs on Reuters4 are reported. From these experiments, it can be seen that the proposed method is relatively robust with respect to these parameters. We also observed similar patterns from experiments on the other datasets.

5. Conclusions

This paper presents a novel machine learning problem – Multiple Instance LEArning with Global Embedding (MILEAGE) for integrating the global feature representations into multiple instance learning. To solve the proposed problem, a large margin method is proposed, which adaptively tunes the weights for the two different feature representations imposed on each bag and trains the classifier. To solve the resulted non-convex non-smooth problem efficiently, an alternative method is employed and the bundle method that explicitly treats the convex and non-convex parts is suggested. Some theoretical analysis, such as the time complexity and generalized error rate, are provided thereafter. The experimental results on both the text and image datasets, as well as the newly proposed application – Insider Threat Detection, clearly demonstrate the advantages of the proposed method. However, since MILEAGE is a newly proposed research problem, some future works are still necessary. In the future, we plan to (1) design some other frameworks to formulate this problem such that both accuracy and efficiency can be further increased; (2) investigate whether some other intuitions, other than the positive ratios in each bag, can be employed to design methods for MILEAGE; (3) extend the current binary classification method to the multi-label case (Zhou & Zhang, 2006). (4) consider how to incorporate the structure information between examples (Zhang et al., 2011) into the framework of MILEAGE. (5) treat the local and global representations in different feature spaces.

Acknowledgments

This work is partially supported by NSF research grants IIS-0746830, CNS- 1012208 and IIS-1017837. This work is also partially supported by the Center for Science of Information (CSoI) under grant agreement CCF-0939370.

References

Andrews, S., Tsochantaridis, I., and Hofmann, T. Support vector machines for multiple-instance learning. In *NIPS*, 2003.

Babenko, Boris, Yang, Ming-Hsuan, and Belongie, Serge. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1619–1632, 2011.

Bergeron, Charles, Moore, Gregory M., Zaretzki, Jed, Breneman, Curt M., and Bennett, Kristin P. Fast bundle algorithm for multiple-instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6):1068–1079, 2012.

Berry, Michael W. and Castellanos, Malu. Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition. 2007.

Chen, Yixin, Bi, Jinbo, and Wang, James Ze. Miles: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, 2006.

Dietterich, T. G., Lathrop, R. H., and Lozano-Perez, T. Solving the multiple instance problem with axis-parallel rectangles. In *Artificial Intelligence*, 1998.

Do, Trinh Minh Tri and Artières, Thierry. Large margin training for hidden markov models with partially observed states. In *ICML*, pp. 34, 2009.

Fu, Zhouyu and Robles-Kelly, Antonio. An instance selection approach to multiple instance learning. In *CVPR*, pp. 911–918, 2009.

Fuduli, Antonio, Gaudioso, Manlio, and Giallombardo, Giovanni. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14(3): 743–756, 2004.

Hare, Warren and Sagastizábal, Claudia A. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.

Hu, Yang, Li, Mingjing, and Yu, Nenghai. Multiple-instance ranking: Learning to rank images for image retrieval. In *CVPR*, 2008.

Joachims, Thorsten. Training linear svms in linear time. In *KDD*, pp. 217–226, 2006.

Joachims, Thorsten, Cristianini, Nello, and Shawe-Taylor, John. Composite kernels for hypertext categorisation. In *ICML*, pp. 250–257, 2001.

Joachims, Thorsten, Finley, Thomas, and Yu, Chun-Nam John. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

Kim, Minyoung and la Torre, Fernando De. Gaussian processes multiple instance learning. In *ICML*, pp. 535–542, 2010.

Manning, Christopher D., Raghavan, Prabhakar, and Schtze, Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Noll, Dominikus. Bundle method for non-convex minimization with inexact subgradients and function values. *Computational and Analytical Mathematics*, 2012.

Rahmani, R. and Goldman, S.A. MISSL: Multiple-instance semi-supervised learning. In *ICML*, 2006.

Ray, Soumya and Craven, Mark. Supervised versus multiple instance learning: an empirical comparison. In *ICML*, pp. 697–704, 2005.

Schramm, Helga and Zowe, Jochem. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121, 1992.

Shawe-Taylor, John and Cristianini, Nello. *Kernel Methods for Pattern Analysis*. MCambridge University Press, 2004.

Smola, Alex J., Vishwanathan, S. V. N., and Le, Quoc V. Bundle methods for machine learning. In *NIPS*, 2007.

Teo, Choon Hui, Vishwanathan, S. V. N., Smola, Alex J., and Le, Quoc V. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11:311–365, 2010.

Wang, Jun and Zucker, Jean-Daniel. Solving multiple-instance problem: A lazy learning approach. In Langley, Pat (ed.), *ICML*, pp. 1119–1125, 2000.

Wu, Ou, Gao, Jun, Hu, Weiming, Li, Bing, and Zhu, Mingliang. Indentifying multi-instance outliers. In *SDM*, pp. 430–441, 2010.

Yuille, A. and Rangarajan, A. The concave-convex procedure. *Neural Computation*, 2003.

Zhang, Dan, Liu, Yan, Si, Luo, Zhang, Jian, and Lawrence, Richard D. Multiple instance learning on structured data. In *NIPS*, pp. 145–153, 2011.

Zhou, Z-H, Sun, Y-Y, and Li, Y-F. Multi-instance learning by treating instances as non i.i.d. samples. In *ICML*, 2009.

Zhou, Zhi-Hua and Zhang, Min-Ling. Multi-instance multi-label learning with application to scene classification. In *NIPS*, pp. 1609–1616, 2006.