# MI2LS: Multi-Instance Learning from Multiple Information Sources

Dan Zhang
Facebook Incorporation
Menlo Park, CA
danzhang2008@gmail.com

Jingrui He
Computer Science Department
Stevens Institute of Technology
Hoboken, NJ
jingrui.he@gmail.com

Richard D. Lawrence
Machine Learning Group
IBM T.J. Watson Research Center
Yorktown Heights, NY
ricklawr@us.ibm.com

## ABSTRACT

In Multiple Instance Learning (MIL), each entity is normally expressed as a set of instances. Most of the current MIL methods only deal with the case when each instance is represented by one type of features. However, in many real world applications, entities are often described from several different information sources/views. For example, when applying MIL to image categorization, the characteristics of each image can be derived from both its RGB features and SIFT features. Previous research work has shown that, in traditional learning methods, leveraging the consistencies between different information sources could improve the classification performance drastically.

Out of a similar motivation, to incorporate the consistencies between different information sources into MIL, we propose a novel research framework – Multi-Instance Learning from Multiple Information Sources (MI$^2$LS). Based on this framework, an algorithm – Fast MI$^2$LS (FMI$^2$LS) is designed, which combines Constraint Concave-Convex Programming (CCCP) method and an adapted Stoachastic Gradient Descent (SGD) method. Some theoretical analysis on the optimality of the adapted SGD method and the generalized error bound of the formulation are given based on the proposed method. Experimental results on document classification and a novel application – Insider Threat Detection (ITD), clearly demonstrate the superior performance of the proposed method over state-of-the-art MIL methods.

## Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning – Knowledge acquisition

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Multi-View Learning, Multi-Instance Learning, Stoachastic Gradient Descent

*KDD'13,* August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

## 1. INTRODUCTION

*Traditional learning* methods normally treat each example as a non-separable entity, and represent the example by one feature vector. However, the semantic meanings of each individual example could vary among its constituent parts, rather than being consistent throughout the whole content. As one variation of traditional learning methods, Multiple Instance Learning (MIL) [11] has been proposed to solve the label ambiguity problem. In particular, in MIL, each example/bag is divided into several different parts/instances. The labels are assigned to the bags, rather than individual instances. In this way, the features for the desired local object in each example will be less likely affected by its irrelevant parts, and therefore the learned model can be more accurate. A lot of work has been done for MIL classification [2, 11, 14, 23, 32, 46] and its variants, such as outlier detection [44], online learning [3], and ranking [19]. These methods have been widely employed in applications such as text mining [2], drug design [11], localized content based image retrieval (LCBIR) [32], human action recognition [1] and market targeting [46].

Most of the current MIL methods focus merely on solving problems where examples are described by only one set of features. However, in many real-world applications, examples are often derived from several different information sources/views, and therefore are represented by multiple sets of features. For example, in webpage classification, each webpage has disparate descriptions such as in-bound, out-bound links and textual content. In image retrieval, each image can be described by different kinds of features, such as RGB features, SIFT features [27], and texture features. Different sets of features normally have different statistical properties. As shown in previous studies in multi-view learning work [5, 12, 22, 25, 35, 39, 50], by leveraging the consistencies between different views, the classification performance can be improved. Therefore, designing a MIL algorithm that incorporates information from multiple sources is also expected to bring in performance improvements.

The existing research in this direction is rare. In [31], the authors did some experiments by using MIL on different views separately and then combined them with equal weights. This method is straightforward. However, it does not consider the consistencies between different views. On the contrary, in this paper, to integrate the consistencies into MIL, a novel framework – Multi-Instance Learning from Multiple Information Sources (MI$^2$LS) is proposed. From the MIL perspective, MI$^2$LS integrates the nature of the multi-view setting into the MIL framework and impose the consistencies among multiple views. From multi-view learning perspective, the new formulation explicitly handles the problem of label ambiguity through modeling different segments of examples. More precisely, the new framework aims at designing classifiers for MIL on individual views and constraining the consistencies between these classifiers simultaneously. Based on the proposed framework, a concrete optimization formulation is suggested. However, the proposed formulation is non-convex and contains too many constraints derived on both the bag and the instance levels. Therefore, to solve the resulting optimization problem, we propose a novel method – Fast MI$^2$LS (FMI$^2$LS), which is a combination of Constrained Concave-Convex Procedure (CCCP) and Stochastic Gradient Descent (SGD). We prove that the proposed method is guaranteed to converge with some derived convergence bounds. Furthermore, the generalized error bound of the proposed method is analyzed. To show the effectiveness and efficiency of the proposed method, in the experiment part, a series of experiments are conducted on two benchmark text datasets, Reuters21578, WebKB, as well as a newly introduced application of MIL – Insider Threat Detection (ITD). In this new application, MIL is employed to find the potential harmful insiders through analyzing their online behaviors, where the features during each time period is modeled as a bag and each bag contains instances derived from daily features. The different views in ITD indicate different types of online behaviors. Experimental results on this application and the two text datasets clearly demonstrate the advantages of our proposed method over state-of-the-art techniques.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 proposes the research problem and presents the proposed algorithm. Some theoretical analysis are given in Section 4. Section 5 presents the experimental results. Section 6 concludes the whole paper.

# 2. RELATED WORKS

## 2.1 Multi-Instance Classification

The concept of MIL was first introduced by Dietterich et al. [11] for predicting musk molecular. Since then, numerous research work has been done in MIL. Roughly speaking, MIL methods can be separated into three groups, (1) the group that is specifically designed to solve MIL [11, 29]; (2) the group that converts MIL to traditional single-instance problems and solve the resulting problem through traditional learning methods [8, 9]. (3) the group that revises traditional single-instance learning methods by imposing MIL constraints [2, 7, 15, 16, 23, 24].

For the first group, APR [11], which encloses positive instances by an axis-parallel rectangle in the feature space, is the first method to solve MIL problems. Later, Maron and Lozano-Pérez proposed Diverse Density (DD) [29, 30], which tries to identify the concept point that resembles positive instance most, and classify unlabeled bags according to the distances between the instances in these bags and this concept point. In [33, 49], the authors accelerated DD method by applying Expectation-Maximization (EM), and proposed EM-DD.

In the second group, DD-SVM [9] picks a set of prototypes among the local solutions from DD method returned by different initializations and then design a large margin classifier based on the bag level features extracted from these selected prototypes. In [8], the authors embedded bags into a feature space spanned by instances, and apply 1-norm SVM to build the bag level classifiers.

Most of the MIL methods fall into the third group. Andrew et al.[2] proposed two different MIL formulations based on SVM [6], i.e., misvm for the instance level classification and MISVM for the bag level classification. Since the MIL formulations are non-convex, Gehler and Chapelle tried to use deterministic annealing and achieved better local solutions [16]. Gärtner et al. [15] put forward a kernel function directly based on bags. Later, Kwok and Cheung [24] advanced their work through proposing a marginalized MIL kernel and converting the MIL from an incomplete data problem to a complete data problem. In [7], the authors revised the loss functions of single-instance SVM and focus more on the positive bags with smaller sizes. To improve the efficiency of misvm and MISVM, bundle method is adapted to solve the non-convex optimization problem [4]. Furthermore, some research work incorporates the MIL constraints into gaussian process [23] and conditional random fields [10]. In [20, 26, 48, 51], the multi-instance multi-label problem has also attracted a lot of attentions, in which the labels are not restricted to be binary, but can be a vector. Moreover, some other variants of MIL are also proposed, such as multi-instance outlier detection [44], multi-instance online learning [3] and multi-instance ranking [19].

The previous research work is reasonable, and solves emerging MIL problems from different perspectives. However, few of them considered the case when examples are derived from multiple information sources, while the previous work on traditional single instance learning methods has demonstrated superior performances of methods that consider the consistencies between different information sources over the ones that do not. Out of this motivation, the proposed framework MI$^2$LS integrates the consistencies between different sources into a unified framework for MIL, and Fast MI$^2$LS is proposed to solve the suggested formulation in an efficient and effective way.

## 2.2 Learning with Multiple Information Sources

In a lot of real-world applications, examples are usually extracted from multiple information sources/views. It has been shown extensively in prior research that utilizing the consistency between the multiple sources/views could achieve better performance [5, 12, 22, 25, 35, 39, 45, 47, 50]. In particular, one of the earliest work in multi-view learning is [5], in which the authors propose the co-training method to solve problems where the examples are described by two distinct views. In [12], the authors build classifiers on different views and constrain the consistencies between different classifiers on each individual view. Moreover, they show that the Rademacher complexity of the function class can also be greatly reduced by regulating the consistencies.

This idea is further exploited in [25], in which the consistency term is incorporated into multi-view semi-supervised learning problems, and it has shown a substantial improvement on the classification performance. Likewise, in [47], the authors introduce the consistency into local learning [43] and design a novel way to define the graph Laplacian. When applied to transfer learning [17, 45], imposing the consistencies between different views also shows superior performances in transferring the knowledge between different domains. Most existing multi-view learning methods are for the single instance settings, while MIL problem naturally exists in real world applications. So, different from the prior work, in this paper, the view consistency constraint is further applied to MIL problems, such that the label ambiguity problem in multi-view learning can be handled in a more principled way.

# 3. THE PROPOSED METHOD

## 3.1 Problem Statement and Notation

Suppose a set of $n$ labeled bags: $\mathcal{D} = \{(\mathbf{B}_i, \mathbf{Y}_i), i = 1, \ldots, n\}$ are available for training, where $\mathbf{B}_i$ represents the $i$-th bag and $\mathbf{Y}_i \in \{1, -1\}$ is its binary label. The bag $\mathbf{B}_i$ consists of a set of instances, and each instance is described by different views. In particular, the $p$-th view of instances in the $i$-th bag $\mathbf{B}_i$ are denoted as $\{\mathbf{B}_{i1}^{(p)}, \ldots, \mathbf{B}_{in_i}^{(p)}\}$, $p = 1, \ldots, M$, and $\mathbf{B}_{ij}^{(p)} \in \mathcal{R}^{d_p}$[1]. $d_p$ is the dimensionality of the $p$-th view. $n_i$ is the number of instances in the $i$-th bag and $M$ is the total number of views. The objective of Multi-Instance Learning from Multiple Information Sources (MI$^2$LS) is to design a function $f : \mathbf{B} \rightarrow \{1, -1\}$ by integrating the consistencies between different views into MIL, such that classification on the unlabeled bags could be accurate.

## 3.2 Formulation

We aim to leverage the instances derived from different information sources (views) and their labels simultaneously. The general framework of MI$^2$LS is as follows:

$$\begin{aligned}\min_{\mathbf{w}^{(p)}} \ & \Omega(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}) + L_c(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}) \\ & + L_a(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}),\end{aligned}$$

where $\Omega(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$ is regularizer that depicts the capacity of the classifiers on different views, $L_c(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$ represents the classification loss on the different views given by the classifiers, $L_a(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$ measures the consistencies of the classifiers on different views based on the corresponding classification outputs. Since in MIL the outputs can be measured on both the instance level and the bag level, $L_a(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$ can also be defined on the bag level, the instance level or on both of the two levels. Through incorporating these three components, the proposed framework ensures that the classification on each individual view should be accurate enough and the output of each individual instance or bag given by the classifiers on different views should be consistent.

Following this framework and considering the case when features are derived from two views without the loss of generality ($M = 2$), there are multiple ways of formulating the three different terms. For the first part, one of the possible options to define the regularizer, which is also the one used in this paper, is $\Omega(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}) = \sum_{p=1}^{2} \|\mathbf{w}^{(p)}\|^2$. The hinge loss can be applied to $L_c(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$ similar to most large marge methods. The $\epsilon$-insensitive loss is used to define $L_a(\mathcal{D}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$[2], which requires the inconsistency between different views of each instance be within $\epsilon$ error bound and penalizes the discrepancy beyond this bound. Then, a concrete formulation can be given as:

$$\begin{aligned}\min_{\mathbf{w}^{(1)},\mathbf{w}^{(2)}} \quad & \frac{1}{2}\sum_{p=1}^{2}\|\mathbf{w}^{(p)}\|^2 + \frac{1}{n}\sum_{p=1}^{2}\sum_{i=1}^{n} C^{(p)}\xi_i^{(p)} + \frac{C}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\eta_{ij} \\ s.t. \quad & \forall i \in \{1, 2, \ldots, n\} \\ & \mathbf{Y}_i \max_{j \in n_i} \mathbf{w}^{(1)T}\mathbf{B}_{ij}^{(1)} \geq 1 - \xi_i^{(1)} \\ & \mathbf{Y}_i \max_{j \in n_i} \mathbf{w}^{(2)T}\mathbf{B}_{ij}^{(2)} \geq 1 - \xi_i^{(2)} \\ & \forall i \in \{1, 2, \ldots, n\}, \forall j \in \{1, 2, \ldots, n_i\} \\ & |\mathbf{w}^{(1)T}\mathbf{B}_{ij}^{(1)} - \mathbf{w}^{(2)T}\mathbf{B}_{ij}^{(2)}| \leq \epsilon + \eta_{ij}, \end{aligned} \tag{1}$$

where $N = \sum_{i=1}^{n} n_i$, $C^{(1)}$, $C^{(2)}$ and $C$ are trade-off parameters tuning the importances on the classification losses on the corresponding views as well as the penalty term that measures the consistencies between different views.

The proposed optimization formulation imposed the view consistency assumption into the framework of MIL in a reasonable way. However, this is a non-convex optimization problem. So, it cannot be solved directly. Moreover, in many real world problems, the numbers of bags and instances are huge, which would result in a large number of constraints and therefore could drastically increase the computational complexity for solving this problem. To deal with this optimization problem efficiently and effectively, a concrete method – Fast MI$^2$LS (FMI$^2$LS) is therefore proposed in the following sections.

## 3.3 Method

For the convenience of computation, without loss of generality, we introduce three concatenated vectors as:

$$\begin{aligned} & \widetilde{\mathbf{w}} = [\mathbf{w}^{(1)T}, \mathbf{w}^{(2)T}]^T, \\ & \widetilde{\mathbf{B}}_{ij}^{(1)} = [\mathbf{B}_{ij}^{(1)T}, \mathbf{0}^{d_2 T}]^T, \widetilde{\mathbf{B}}_{ij}^{(2)} = [\mathbf{0}^{d_1 T}, \mathbf{B}_{ij}^{(2)T}]^T, \end{aligned}$$

where $\mathbf{0}^{d_p}$ is a $1 \times d_p$ zero vector. After this transformation, it is clear that $\mathbf{w}^{(p)T}\widetilde{\mathbf{B}}_{ij}^{(p)} = \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(p)}$. Then, problem (1) can be converted to the following form:

$$\begin{aligned}\min_{\widetilde{\mathbf{w}}} \quad & \frac{1}{2}\|\widetilde{\mathbf{w}}\|^2 + \frac{1}{n}\sum_{p=1}^{2}\sum_{i=1}^{n} C^{(p)}\xi_i^{(p)} + \frac{C}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\eta_{ij} \\ s.t. \quad & \forall i \in \{1, 2, \ldots, n\} \\ & \mathbf{Y}_i \max_{j \in n_i} \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)} \geq 1 - \xi_i^{(1)} \\ & \mathbf{Y}_i \max_{j \in n_i} \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)} \geq 1 - \xi_i^{(2)} \\ & \forall i \in \{1, 2, \ldots, n\}, \forall j \in \{1, 2, \ldots, n_i\} \\ & \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)} \leq \epsilon + \eta_{ij} \\ & \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)} \geq -\epsilon - \eta_{ij}. \end{aligned} \tag{2}$$

Compared with problem (1), although this form is simplified, it is still non-convex and contains too many constraints. There are multiple ways of handling the non-convex optimization problems, such as Constrained Concave-Convex Procedure (CCCP) [41], adapted bundle method [13] and deterministic annealing [16]. Due to the popularity of CCCP, we use this method to decompose this non-convex problem into a series of convex sub-problems and focus on the resulting convex subproblems. Furthermore, to reduce the time complexity on solving these subproblems, Stochastic Gradient Descent (SGD) [37] method is adapted, such that the algorithm can find a local optimal solution in linear scale.

[1] In MI$^2$LS, the instances on different views could be derived from different partition ways and the numbers of instances in the same bag could be different on different views. Here, we do not consider this case out of simplicity. As we shall see later, the proposed framework could handle this situation by imposing consistencies on the bag level.

[2] In the proposed method, for simplicity, we only consider the case when the consistency is defined on the instance level. If the consistency is defined on the bag level, then the last constraint of problem (1) can be re-written to restrict the differences between the outputs of each bag on different views in a similar way. The resulting optimization problem can be solved using a similar method as the one proposed in this paper. If the consistency is defined on both of the two levels, the constraint can be considered as a combination of the bag and instance level consistencies.

## 3.4 CCCP with Stochastic Gradient Descent

Given a starting point $\widetilde{\mathbf{w}}^{(0)}$[3], CCCP iteratively computes $\widetilde{\mathbf{w}}^{(t)}$ from $\widetilde{\mathbf{w}}^{(t-1)}$ by replacing $\max_{j\in n_i} \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)}$ and $\max_{j\in n_i} \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)}$ with their first order Taylor expansions at $\widetilde{\mathbf{w}}^{(t-1)}$. More precisely, for the $t$-th iteration of CCCP, the derived subproblem for solving problem (2) is:

$$\begin{aligned}
\min_{\widetilde{\mathbf{w}}} \quad & \frac{1}{2}\|\widetilde{\mathbf{w}}\|^2 + \frac{1}{n}\sum_{p=1}^{2}\sum_{i=1}^{n} C^{(p)}\xi_i^{(p)} + \frac{C}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\eta_{ij} \\
s.t. \quad & \forall i \in \{1,2,\ldots,n\} \\
& \mathbf{Y}_i\widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij_1^*}^{(1)} \geq 1-\xi_i^{(1)} \\
& \mathbf{Y}_i\widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij_2^*}^{(2)} \geq 1-\xi_i^{(2)} \\
& \forall i \in \{1,2,\ldots,n\}, \forall j \in \{1,2,\ldots,n_i\} \\
& \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)} \leq \epsilon + \eta_{ij} \\
& \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)} \geq -\epsilon - \eta_{ij} \qquad (3)
\end{aligned}$$

where $j_p^* = \arg\max_j \widetilde{\mathbf{w}}^{(t-1)T}\widetilde{\mathbf{B}}_{ij}^{(p)}$, and represents the most positive instance for the $i$-th bag on $p$-th view. Through solving a series of subproblems derived from CCCP, the method is guaranteed to converge to a local optimal solution of problem (2).

The resulting subproblem is convex. However, the cost of directly solving this problem is non-trivial, especially when the numbers of bags, instances, as well as the resulting constraints for the optimization problem are large. A lot of research work, such as bundle method [21, 40] and SGD method, has been proposed to improve the efficiency of similar optimization problems. In this paper, due to the superior performance, SGD is employed. Different from the traditional SGD method, in problem (3), we have two different sets of constraints, i.e., the ones on the bags and the ones on instances. The algorithm receives several parameters, i.e., $S$ - the number of SGD iterations to perform; $k_1$ and $k_2$ ($k_1 << n$, $k_2 << N$) - the number of bags and instances to use for approximating the sub-gradients. At the beginning of SGD algorithm for the $t$-th CCCP iteration, we set $\widetilde{\mathbf{w}}^{(t_0)}$ to be $\widetilde{\mathbf{w}}_{\alpha}^{(t-1)_S}$ [4], whose norm is at most $\sqrt{C^{(1)}+C^{(2)}}$. Here, the subscript $\alpha$ means that the output of SGD for the $t$-th CCCP iteration is an averaged result of the last corresponding $\alpha S$ SGD iterations. The averaged result is adopted here because of the superior performance as shown in [34]. For the $s$-th iteration of the SGD algorithm, we randomly pick a set of bags $A_s \in \{1,\ldots,n\}$, and another set of instances $B_s$ from all of the instances (as indicated by $\overline{A}_s$) in selected bags. By doing so, the computational cost can be reduced on both the bag level and the instance level. More precisely, during each SGD iteration, we replace problem (3) with an approximated convex sub-problem as follows:

$$\begin{aligned}
&\min_{\widetilde{\mathbf{w}}} f(\widetilde{\mathbf{w}}, A_s, B_s) \qquad (4) \\
&= \frac{1}{2}\|\widetilde{\mathbf{w}}\|^2 + \frac{1}{k_1}\sum_{p=1}^{2}\sum_{i=1}^{k_1} C^{(p)}\max\{0, 1-\mathbf{y}_i\widetilde{\mathbf{w}}^T\mathbf{x}_i^{(p)}\} \\
&+ \frac{C}{k_2}\sum_{i=1}^{k_2}\max\left\{\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)} - \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)} - \epsilon,\ \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)} - \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)} - \epsilon, 0\right\},
\end{aligned}$$

where, $(\mathbf{x}_i^{(p)}, \mathbf{y}_i)$ represents the instance whose output is the largest in the corresponding bag given by the classifier $\widetilde{\mathbf{w}}^{t_0}$ , i.e., the instances indicated by $j_p^*$ in Eq.(3), and its corresponding label from the selected bags in $A_s$[5]. $\mathbf{z}_i \in B_s$ represents the instances sampled from all of the instances in selected bags, i.e., $\overline{A}_s$. It is clear that the subgradient of $f(\widetilde{\mathbf{w}}, A_s, B_s)$ can be calculated as:

$$\begin{aligned}
\frac{\partial f(\widetilde{\mathbf{w}}, A_s, B_s)}{\partial\widetilde{\mathbf{w}}} =& \widetilde{\mathbf{w}} - \frac{1}{k_1}\sum_{p=1}^{2}\sum_{i=1}^{k_1} C^{(p)} I_{i1}^{(p)}\mathbf{y}_i\mathbf{x}_i^{(p)} \\
&+ \frac{C}{k_2}\sum_{i=1}^{k_2}(I_{i2}-I_{i3})(\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}), \qquad (5)
\end{aligned}$$

where $I_{i1}^{(p)}$, $I_{i2}$, $I_{i3}$ are indicator functions. $I_{i1}^{(p)}$ equals 1, if $\mathbf{y}_i\widetilde{\mathbf{w}}^T\mathbf{x}_i^{(p)} < 1$, and otherwise 0; $I_{i2}$ equals 1 if $\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)} - \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)} > \epsilon$ and otherwise 0; $I_{i3}$ equals 1 if $\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)} - \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)} > \epsilon$ and otherwise 0. By setting the step length to be $\eta_s = \frac{1}{s}$, the updating scheme can be written as $\widetilde{\mathbf{w}}^{t_{s+1}} = \widetilde{\mathbf{w}}^{t_s} - \eta_s \frac{\partial f(\widetilde{\mathbf{w}}, A_s, B_s)}{\partial\widetilde{\mathbf{w}}} |_{\widetilde{\mathbf{w}}=\widetilde{\mathbf{w}}^{t_s}}$. $\widetilde{\mathbf{w}}^{t_{s+1}}$ will then be projected to the set $\{\|\widetilde{\mathbf{w}}\| \leq \sqrt{C^{(1)}+C^{(2)}}\}$. Here, $\sqrt{C^{(1)}+C^{(2)}}$ is radius of the ball that the optimal solution of (4) should fall into, as shown in the later section. The final output is averaged from $\widetilde{\mathbf{w}}^{t_{(1-\alpha)S}}$ to $\widetilde{\mathbf{w}}^{t_S}$ as: $\widetilde{\mathbf{w}}_{\alpha}^{t_S} = \frac{\widetilde{\mathbf{w}}^{t_{(1-\alpha)S}} + \ldots + \widetilde{\mathbf{w}}^{t_S}}{\alpha S}$ for some constant $\alpha \in (0,1)$. Based on the above derivation, the whole algorithm can be summarized in Table 1.

# 4. THEORETICAL ANALYSIS

In this section, some important properties of the proposed method, such as the optimality and generalized error rate, will be analyzed.

## 4.1 Optimality

It has already been shown in previous work that the CCCP [41] will converge asymptotically. During each CCCP iteration, SGD is used for solving the resulting convex sub-problem. In this section, we will investigate some important properties of the adapted SGD method, such as the bound of the optimal solution and the difference between the objective function values of $\widetilde{\mathbf{w}}_{\alpha}^{t_S}$ and that of $\widetilde{\mathbf{w}}^{t*}$, where $\widetilde{\mathbf{w}}^{t*}$ refers to the optimal value for the $t$-th iteration.

**Theorem 1:** Suppose $G(\widetilde{\mathbf{w}}, A_s, B_s) = \frac{1}{k_1}\sum_{p=1}^{2}\sum_{i=1}^{k_1} C^{(p)}$ $\max\{0, 1-\mathbf{y}_i\widetilde{\mathbf{w}}^T\mathbf{x}_i^{(p)}\} + \frac{C}{k_2}\sum_{i=1}^{k_2}\max\{\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)} - \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)} - \epsilon, \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)} - \widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)} - \epsilon, 0\}$. Then, $\|\partial G(\mathbf{w}, A_s, B_s)\|^2 \leq H^2$, where, $H^2 = C^{(1)2}U^{(1)2} + C^{(2)2}U^{(2)2} + C^2\max\{U^{(1)2}, U^{(2)2}\} + 2C^{(1)}C^{(2)}U^{(1)}U^{(2)} + 2C^{(1)}CU^{(1)2} + 2C^{(2)}CU^{(2)2}$, $U^{(1)} = \max_{i,j}\{\|\mathbf{B}_{ij}^{(1)}\|\}$, $U^{(2)} = \max_{i,j}\{\|\mathbf{B}_{ij}^{(2)}\|\}$.

**Proof:** Please check the Appendix.

**Theorem 2:** The optimal solution of problem (3) should fall within the ball of $\sqrt{C^{(1)}+C^{(2)}}$.

**Proof:** Please check the Appendix.

Theorem 2 justifies the reason why during each SGD iteration, at step 11 of Table 1, the solution will be regulated within the ball $\sqrt{C^{(1)}+C^{(2)}}$, since the optimal solution is guaranteed to be falling within this ball.

**Theorem 3:** For the $s$-th SGD iteration, the following inequality holds [6]:

$$\frac{1}{S}\sum_{s=1}^{S} f_s(\widetilde{\mathbf{w}}^{t_s}) \leq \frac{1}{S}\sum_{s=1}^{S} f_s(\widetilde{\mathbf{w}}^{t*}) + \frac{(H^2 + C^{(1)} + C^{(2)})(1 + lnS)}{S}$$

[3] $\widetilde{\mathbf{w}}^{(t)}$ represents the result from the $t$-th CCCP iteration.

[4] Here, the superscript $t_s$ means the $s$-th SGD iteration for the $t$-th CCCP iteration.

[5] To avoid confusion, please note that $\mathbf{y}_i$ indicates the label for the $i$-th selected bag from $A_s$, while $\mathbf{Y}_i$ in the previous formulations refers to the label for the $i$-th bag in the whole dataset.

[6] $f_s(\widetilde{\mathbf{w}}^{t_s})$ here refers to $f(\widetilde{\mathbf{w}}^{t_s}, A_s, B_s)$ in problem (4).

**Table 1: The description of FMI$^2$LS**

**Input:** 1. Labeled bags: $\{(\mathbf{B}_i, Y_i), i = 1, 2, \cdots, n\}$; 2. parameters: trade-off parameters $C^{(1)}$, $C^{(2)}$ and $C$; subsample sizes $k_1$ for bags and $k_2$ for instances; SGD iterations $S$; averaging constant $\alpha$.
**Output:** The classifier $\widetilde{\mathbf{w}}_{\alpha}^{t_S}$.

```
CCCP Iterations:
Initialize $\widetilde{\mathbf{w}}^0$, $t = 0$.
repeat
  Derive problem (3).
  Stochastic Gradient Descent Iterations:
    for $s = 1, \dots, S$
      Choose $A_s \in \mathcal{D}$, where $|A_s| = k_1$.
      Set $A_s^+ = \{(\mathbf{x}_i^{(p)}, \mathbf{y}_i) \in A_s : \mathbf{y}_i \langle \widetilde{\mathbf{w}}^{t_{s-1}}, \mathbf{x}_i^{(p)} \rangle \leq 1\}$.
      Choose $B_s \in \overline{A}_s$, where $|B_s| = k_2$.
      Set $B_s^+ = \{\mathbf{z}_i \in B_s : \max\left\{(\widetilde{\mathbf{w}}^{t_{s-1}})^T \mathbf{z}_i^{(1)} - (\widetilde{\mathbf{w}}^{t_{s-1}})^T \mathbf{z}_i^{(2)} - \epsilon, (\widetilde{\mathbf{w}}^{t_{s-1}})^T \mathbf{z}_i^{(2)} - (\widetilde{\mathbf{w}}^{t_{s-1}})^T \mathbf{z}_i^{(1)} - \epsilon\right\} > 0\}$
      Calculate $\frac{\partial f(\widetilde{\mathbf{w}}, A_s, B_s)}{\partial \widetilde{\mathbf{w}}} |_{\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}}^{t_{s-1}}}$ according to Eq.(5).
      Calculate $\widetilde{\mathbf{w}}^{t_s} = \widetilde{\mathbf{w}}^{t_{s-1}} - \frac{1}{s} \frac{\partial f(\widetilde{\mathbf{w}}, A_s, B_s)}{\partial \widetilde{\mathbf{w}}} |_{\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}}^{t_{s-1}}}$.
      Update $\widetilde{\mathbf{w}}^{t_s} = \min\{1, \frac{\sqrt{C^{(1)} + C^{(2)}}}{\|\widetilde{\mathbf{w}}^{t_s}\|}\} \widetilde{\mathbf{w}}^{t_s}$.
    end for
  $t = t + 1$.
  $\widetilde{\mathbf{w}}^{(t_0)} = \widetilde{\mathbf{w}}_{\alpha}^{(t-1)_S}$.
until convergence
$\widetilde{\mathbf{w}}_{\alpha}^{t_S} = (\widetilde{\mathbf{w}}^{t_{(1-\alpha)S}} + \dots + \widetilde{\mathbf{w}}^{t_S}) / \alpha S$.
```

**Proof:** $\|\frac{\partial f_s(\widetilde{\mathbf{w}})}{\partial \widetilde{\mathbf{w}}}\|^2 \leq (\|\widetilde{\mathbf{w}}\| + H)^2 \leq (\sqrt{C^{(1)} + C^{(2)}} + H)^2$. By plugging this result to Corollary 1 of [36], we can get:

$$\begin{aligned} \frac{1}{S} \sum_{s=1}^{S} f_s(\widetilde{\mathbf{w}}^{t_s}) \leq & \frac{1}{S} \sum_{s=1}^{S} f_s(\widetilde{\mathbf{w}}^{t_*}) + \frac{(H + \sqrt{C^{(1)} + C^{(2)}})^2 (1 + lnS)}{2S} \\ \leq & \frac{1}{S} \sum_{s=1}^{S} f_s(\widetilde{\mathbf{w}}^{t_*}) + \frac{(H^2 + C^{(1)} + C^{(2)})(1 + lnS)}{S} \end{aligned}$$

**Theorem 4:** With probability over the choices of $(A_1, \dots, A_S)$ and $(B_1, \dots, B_S)$, we have that:

$$E[F(\widetilde{\mathbf{w}}_{\alpha}^{t_S}) - F(\widetilde{\mathbf{w}}^{t_*})] \leq \frac{2 + \frac{5}{2} \log(\frac{1}{1-\alpha})}{\alpha} \frac{(\sqrt{C^{(1)} + C^{(2)}} + H)^2}{S},$$

where $F(*)$ is the objective function in problem (3).
**Proof:** $\|\frac{\partial f_s(\widetilde{\mathbf{w}})}{\partial \widetilde{\mathbf{w}}}\|^2 \leq (\|\widetilde{\mathbf{w}}\| + H)^2 \leq (\sqrt{C^{(1)} + C^{(2)}} + H)^2$. By plugging this into Theorem 5 of [34], we can get this conclusion.

## 4.2 Generalized Error Bound

In this section, we consider the class of functions $\mathcal{F}_{C^{(1)}+C^{(2)},D} = \{g | g : \widetilde{\mathbf{B}}_* \longmapsto \frac{1}{2}(\max_j \widetilde{\mathbf{w}}^T \widetilde{\mathbf{B}}_{*j}^{(1)} + \max_j \widetilde{\mathbf{w}}^T \widetilde{\mathbf{B}}_{*j}^{(2)})\}$ such that $\|\widetilde{\mathbf{w}}\|^2 \leq C^{(1)} + C^{(2)}$, and with probability of at least $1 - \delta$,

$$\begin{aligned} & |\widetilde{\mathbf{w}}^T \widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{w}}^T \widetilde{\mathbf{B}}_{ij}^{(2)}| \leq \epsilon + E(\eta_{ij}) \leq D \\ \Rightarrow & \widetilde{\mathbf{w}}^T (\widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{B}}_{ij}^{(2)})^T (\widetilde{\mathbf{B}}_{ij}^{(1)} - \widetilde{\mathbf{B}}_{ij}^{(2)}) \widetilde{\mathbf{w}} \leq D^2 \\ \Rightarrow & \|\widetilde{\mathbf{w}}\|^2 \leq \frac{D^2}{\min_{ij}((\mathbf{B}_{ij}^{(1)})^2 + (\mathbf{B}_{ij}^{(2)})^2)} \\ \Rightarrow & \|\widetilde{\mathbf{w}}\|^2 \leq E^2, \end{aligned} \tag{6}$$

where $E \triangleq D / \sqrt{\min_{ij}((\mathbf{B}_{ij}^{(1)})^2 + (\mathbf{B}_{ij}^{(2)})^2)}$.

**Theorem 5:** The empirical Rademacher complexity of the functional space $\mathcal{F}_{C^{(1)}+C^{(2)},D}$ on $\mathcal{D} = \{(\mathbf{B}_i, Y_i), i = 1, \dots, n\}$ is upper bounded by: $\hat{\mathcal{P}}_n(\mathcal{F}_{C^{(1)}+C^{(2)},D}) = \frac{\min\{\sqrt{C^{(1)}+C^{(2)}}, E\}}{n} \times$ $(\max_{\rho_{ij} \geq 0, \rho_i^T \mathbf{1} = 1} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_{ij} K(\mathbf{B}_{ij}^{(1)}, \mathbf{B}_{ij}^{(1)})}$ $+ \max_{\rho_{ij} \geq 0, \rho_i^T \mathbf{1} = 1} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_{ij} K(\mathbf{B}_{ij}^{(2)}, \mathbf{B}_{ij}^{(2)})})$.

**Proof:** Please check the Appendix.

**Theorem 6:** Fix $\kappa \in (0, 1)$. Then, with probability at least $1 - \kappa$, every $g \in \mathcal{F}_{C^{(1)}+C^{(2)},D}$ satisfies: $P(\mathbf{Y}_* \neq sign(g(\widetilde{\mathbf{B}}_*))) \leq$ $\frac{1}{n \sum_{i=1}^{2} C^{(p)}} \sum_{p=1}^{2} \sum_{i=1}^{n} C^{(p)} \max\{0, 1 - \mathbf{Y}_i g(\widetilde{\mathbf{B}}_i^{(p)})\}$ $+ \hat{\mathcal{P}}_n(\mathcal{F}_{C^{(1)}+C^{(2)},D}) + 3\sqrt{\frac{ln(2/\kappa)}{2n}}$.

**Proof:** This result can be got by applying Theorem 5 to Theorem 4.9 in [38].

# 5. EXPERIMENTS

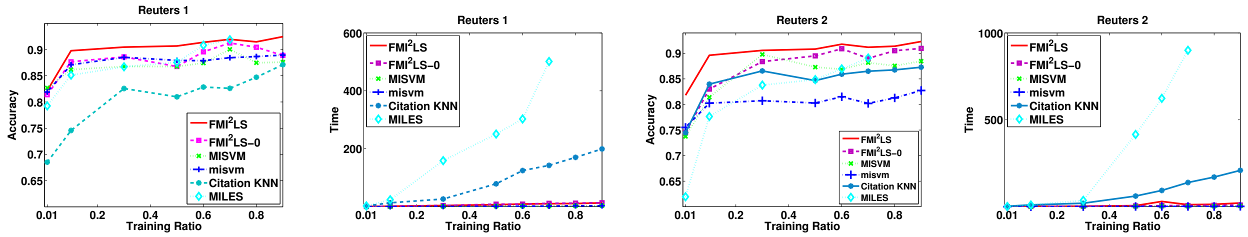
In this section, an extensive set of experiments on document classification and a novel application – insider threat detection is presented to demonstrate the effectiveness and efficiencies of the proposed method.

## 5.1 Datasets

### 5.1.1 Reuters21578

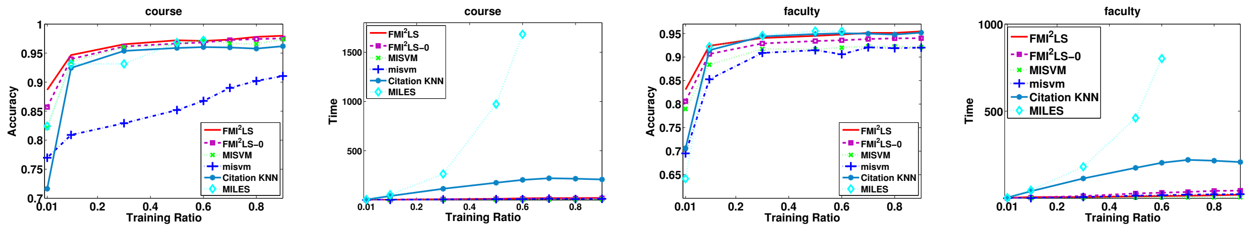
Reuters21578[7] is a benchmark dataset from Reuters newswire in 1987. It has 135 categories, with 21578 documents. We pick documents from 2 sub-categories as the positive examples. The same amount of documents from the remaining dataset are randomly picked as negative ones. In document classification, if a document belongs to a specific category, it is highly possible that not every passage of this document is related to this category. So, it could be better modeled as a MIL problem. More specifically, similar to [2], we treat each document as a bag and use the different

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



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

**Figure 1: Performances Comparisons on Reuters. Some of the experiment results of MILES cannot be reported due to the time complexity issue as stated in Experiment section.**



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

**Figure 2: Performances Comparisons on WebKB. Some of the experiment results of MILES cannot be reported due to the time complexity issue as stated in Experiment section.**

fixed-length passages as instances. For each of the sub-dataset, after removing the stop words and stemming, tf-idf [28] features are extracted and processed by PCA for one information source, and we use the hidden topics information obtained from Probabilistic Latent Semantic Analysis (PLSA)[8] of the binary word features as another one. For a detailed description of these two datasets, please refer to Table 2.

### 5.1.2 WebKB

WebKB[9] is also a benchmark dataset for document classification, which contains webpages from computer science departments in around four different universities. There are seven categories in this dataset, i.e., *student*, *faculty*, *staff*, *department*, *course*, *project* and *other*, with 8280 webpages in this dataset. The two most frequently appeared categories, i.e., course, and faculty, are used for classification, where each sub-dataset contains all of the webpages/bags from one of the two categories, and the same number of the negative bags randomly sampled from the remaining six categories in WebKB. We use the same way as we do for Reusters21578 to extract features from different views and model bags and instances. The detailed description of the two sub-datasets is summarized in Table 2.

| Dataset | # Features View1 | # Features View2 | # Bags | #Instances |
|---|---|---|---|---|
| Reuters1 | 528 | 528 | 1268 | 2367 |
| Reuters2 | 528 | 528 | 1256 | 2145 |
| Course | 320 | 320 | 1348 | 3528 |
| Faculty | 320 | 320 | 1590 | 4248 |
| ITD | 17 | 12 | 1166 | 32235 |

**Table 2: The detailed description of the datasets**

### 5.1.3 Insider Threat Detection (ITD)

We obtained this real dataset from a big IT company. ITD is a project that is devoted to identify the potential harmful insiders through analyzing their online activities, such as sending emails, login, logout, downloaded files, etc. In this project, 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 these labelings, each online activity is quantified as a feature value. However, it is highly possible that a person may not do malicious things on each single day during the period in which he is marked as guilty. 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 have done 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 different activities are quantified into numeric features. These features are further divided into two groups according to the nature of the corresponding behaviors i.e, the group that describes his social behaviors such as sending emails and interacting with friends on social media websites, and the group that depicts things he did by himself, such as logging in and out of a computer system. The whole dataset contains 1000 negative bags and 166 positive bags, where each instance is represented by two different views derived from the two feature groups as described above. Please refer to Table 2 for details on the size of the dataset.

## 5.2 Evaluation Metric

In Reuters21578 and WebKB, since the positive and negative classes are relatively balanced, we use the classification accuracy as the measurement criteria. But for ITD dataset, the number of positive bags is far less than that of the negative ones. So, F1 score for the top 20 returned results is used here for measurement. In particular, F1 score is defined as $F1@20 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$, where, $Precision$ and $Recall$ are measured for the top 20 results.

[8] Actually, PLSA[18] can be considered as a dimensionality reduction method, which maps the documents into some fixed number of hidden topics. The topic distribution for each document can be used as low dimensional features.

[9] http://www.cs.cmu.edu/∼webkb/

## 5.3 Comparison Methods

We compare the proposed method with several state-of-the-art methods. MISVM and misvm [2] are MIL methods based on SVM. The difference between MISVM and midvm is that during each iteration, to update the classifiers, MISVM tries to find a witness for each bag, while misvm assigns pseudo labels to all of the instances. MILES [8] tries to use a single vector to represent each bag through mapping these bags on a learned space. Citation KNN [42] adapts KNN to multi-instance by considering two different kinds of neighborhood relationships. These baseline methods could not be used to solve the multiple view problem directly. So, we concatenate the features in different views together and treat them as from one information source. To demonstrate the benefits of ensuring the consistencies between different views without concatenating the features, we also conduct experiments by setting $C$ to be 0 (FMI$^2$LS-0). It is clear that the formulation of the experiments proposed in [31] can be considered as a special case of FMI$^2$LS-0. For the proposed method, $k_1$ is chosen as $10\%$ of the number of bags, while $k_2$ is $50\%$ of the instances in sampled bags. $\alpha$ is set to be $0.2$. The number of SGD iterations is set to be 30. By using 5 fold cross validation, $C^{(1)}$ and $C^{(2)}$ are searched through the grid $2^{[-5:1:7]}$, $C$ is searched though $2^{[-3:1:5]}$. The parameters of the baseline methods are also tuned similarly.

## 5.4 Results and Analysis

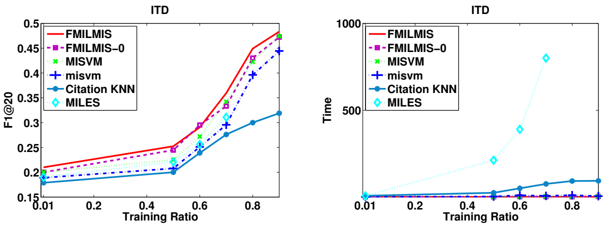
The experiments are conducted by specifying a specific ratio of each dataset for training and keeping the rest for testing. The average results of 20 independent experiments on the three datasets with different training rations are shown in Fig.1, Fig.2 and Fig.3.

From these experimental results, we can see that the proposed method performs better than the other baseline methods in most cases. It is clear that considering the consistencies of examples on different views in MIL could significantly improve the classification performance. The time complexity of the proposed method is also very low, compared with the baseline methods. This is due to the fact that SGD could significantly reduce the time complexity. When compared with FMI$^2$LS-0, it can be concluded that the time complexity of FMI$^2$LS-0 is similar to that of the proposed method. But the performance of FMI$^2$LS-0 is inferior. It further demonstrates the advantages of the proposed method by introducing the consistencies between different views.

For MISVM and misvm, both of these two methods are traditional MIL methods. Their performances are good in terms of both the classification performance and time complexity. However, since these two methods do not consider the different characteristics from multiple information sources, and merely concatenate the different features by using only one feature vector, their performances are inferior to that of the proposed one.

Citation KNN is an adaption of nearest neighbor method. More specifically, it defines two different types of neighbors when measuring the similarities between two bags. It can be seen from the experiments that one of the major drawbacks for this method is that its time complexity is too high because it needs to calculate the distances between test bags and training bags each time. Since it does not consider the consistencies between different views either, contenting the features on different views cannot bring in much additional benefits.

In MILES, during the training phase, the instances in training bags are mapped to a space spanned by the instances in positive bags, and then the most relevant examples are selected through one norm SVM. The method could capture the most important instances in an optimized way. However, the major issue is that its time complexity could be extremely high when the number of instances in positive bags is large. This drawback could potentially hinder its uses in practical applications. In our experiments, this time complexity issue is also very evident. Some experimental results for MILES cannot be acquired due to the extremely large amount of training time. The performance of MILES is very competitive, compared with the other baseline methods. However, it is clear that, from these experiments, its performance cannot exceed the proposed method either.



(a) (b)

**Figure 3: Performances Comparisons on ITD. F1 score for the top 20 returned results is used here due to the imbalance of this dataset. Some of the experiment results of MILES cannot be reported due to the time complexity issue as stated in Experiment section.**

# 6. CONCLUSIONS

In this paper, we investigate an interesting but rarely studied problem – Multi-Instance Learning from Multiple Information Sources (MI$^2$LS). To solve this problem, a general framework is proposed to incorporate the consistencies between different information sources/views into Multi-Instance Learning (MIL). Based on the proposed framework, a concrete method, FMI$^2$LS (Fast MI$^2$-LS) is designed. In particular, the proposed method integrates Constrained Concave-Convex Programming (CCCP) method with an adapted Stoachastic Gradient Descent (SGD) method to solve the non-convex optimization problem in an efficient way. Some important properties of the proposed method are analyzed thereafter. Experimental results on different applications, i.e., document classification and the newly proposed application – Insider Threat Detection (ITD), clearly demonstrate the superior performance of the proposed method against several other state-of-the-art MIL techniques on both efficiency and effectiveness. Based on the proposed method, in the future, we plan to extend the current work in the following ways: (1) In this paper, we didn't tune the weights of different views in the final classifier for simplicity. However, it is often the case that the data quality on different views could be different. We plan to design a method to adaptively tune the weights of different views under the current framework. (2) Due to the nature of MIL, we can define different kinds of consistencies between views, i.e., on the instance level, the bag level, and the mixture of bag and instance level. It is an interesting topic to further investigate which one works better.

## References

[1] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):288–303, 2010.

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

[3] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1619–1632, 2011.

[4] C. Bergeron, G. M. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett. Fast bundle algorithm for multiple-instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6):1068–1079, 2012.

[5] A. Blum and T. M. Mitchell. Combining labeled and unlabeled sata with co-training. In *COLT*, pages 92–100, 1998.

[6] B.Scholkopf and A.Smola. *Learning with Kernels*. MITPress, Cambridge, MA, 2002.

[7] R. C. Bunescu and R. J. Mooney. Multiple instance learning for sparse positive bags. In *ICML*, pages 105–112, 2007.

[8] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, 2006.

[9] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[10] T. Deselaers and V. Ferrari. A conditional random field for multiple-instance learning. In *ICML*, pages 287–294, 2010.

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

[12] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM-2K, theory and practice. *NIPS*, 18:355, 2006.

[13] A. Fuduli, M. Gaudioso, and G. Giallombardo. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14(3):743–756, 2004.

[14] T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola. Multi–instance kernels. In *ICML*, 2002.

[15] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[16] P. V. Gehler and O. Chapelle. Deterministic annealing for multiple-instance learning. *Journal of Machine Learning Research - Proceedings Track*, 2:123–130, 2007.

[17] J. He and R. Lawrence. A graphbased framework for multi-task multi-view learning. In *ICML*, pages 25–32, 2011.

[18] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[19] Y. Hu, M. Li, and N. Yu. Multiple-instance ranking: Learning to rank images for image retrieval. In *CVPR*, 2008.

[20] R. Jin, S. Wang, and Z.-H. Zhou. Learning a distance metric from multi-instance multi-label data. In *CVPR*, pages 896–902, 2009.

[21] T. Joachims. Training linear svms in linear time. In *KDD*, pages 217–226, 2006.

[22] T. Joachims and N. Cristianini. Composite kernels for hypertext categorisation. In *ICML*, pages 250–257, 2001.

[23] M. Kim and F. D. la Torre. Gaussian processes multiple instance learning. In *ICML*, pages 535–542, 2010.

[24] J. T. Kwok and P.-M. Cheung. Marginalized multi-instance kernels. In *IJCAI*, pages 901–906, 2007.

[25] G. Li, S. C. H. Hoi, and K. Chang. Two-view transductive support vector machines. In *SDM*, pages 235–244, 2010.

[26] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 9(1):98–112, 2012.

[27] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[28] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[29] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, 1997.

[30] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, pages 341–349, 1998.

[31] M. Mayo and E. Frank. Experiments with multi-view multi-instance learning for supervised image classification. In *Proc 26th International Conference Image and Vision Computing New Zealand*, Auckland, New Zealand, pages 363–369, 2011.

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

[33] R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts. Localized content-based image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1902–1912, 2008.

[34] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stoachastic optimization. In *ICML*, 2012.

[35] D. Rosenberg, V. Sindhwani, P. Bartlett, and P. Niyogi. A Kernel for Semi-Supervised Learning With Multi-View Point Cloud Regularization. *IEEE Signal Processing Magazine*, 2009.

[36] S. Shalev-shwartz and Y. Singer. Logarithmic regret algorithms for strongly convex repeated games. In *The Hebrew University*, 2007.

[37] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Math. Program.*, 127(1):3–30, 2011.

[38] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. MCambridge University Press, 2004.

[39] V. Sindhwani and P. Niyogi. A co-regularized approach to semi-supervised learning with multiple views. In *ICML Workshop on Learning with Multiple Views*, 2005.

[40] C. H. Teo, S. V. N. Vishwanathan, A. J. Smola, and Q. V. Le. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11:311–365, 2010.

[41] A. S. Vishwanathan, A. J. Smola, and S. V. N. Vishwanathan. Kernel methods for missing variables. In *AISTATS*, pages 325–332, 2005.

[42] J. Wang, et Jean-Daniel Zucker, and J. daniel Zucker. Solving the multiple-instance problem: A lazy learning approach. In *ICML*, pages 1119–1125, 2000.

[43] M. Wu and B. Schölkopf. A local learning approach for clustering. In *NIPS*, pages 1529–1536, 2006.

[44] O. Wu, J. Gao, W. Hu, B. Li, and M. Zhu. Indentifying multi-instance outliers. In *SDM*, pages 430–441, 2010.

[45] D. Zhang, J. He, Y. Liu, L. Si, and R. D. Lawrence. Multi-view transfer learning with a large margin approach. In *KDD*, pages 1208–1216, 2011.

[46] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence. Multiple instance learning on structured data. In *NIPS*, pages 145–153, 2011.

[47] D. Zhang, F. Wang, C. Zhang, and T. Li. Multi-view local learning. In *AAAI*, pages 752–757, 2008.

[48] M.-L. Zhang and Z.-H. Zhou. M3miml: A maximum margin method for multi-instance multi-label learning. In *ICDM*, pages 688–697, 2008.

[49] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *NIPS*, pages 1073–1080, 2001.

[50] T. Zhang, A. Popescul, and B. Dom. Linear prediction models with graph regularization for web-page categorization. In *SIGKDD*, pages 821–826, 2006.

[51] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS*, pages 1609–1616, 2006.

# APPENDIX

**Proof of Theorem 1:** To prove this theorem, we suppose $\iota_i = \frac{\partial \max\{0,1-\mathbf{y}_i\widetilde{\mathbf{w}}^T\mathbf{x}_i^{(1)}\}}{\partial \widetilde{\mathbf{w}}}$, $\kappa_i = \frac{\partial \max\{0,1-\mathbf{y}_i\widetilde{\mathbf{w}}^T\mathbf{x}_i^{(2)}\}}{\partial \widetilde{\mathbf{w}}}$, $\upsilon_i = \frac{\partial \max\{\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)}-\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)}-\epsilon,\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(2)}-\widetilde{\mathbf{w}}^T\mathbf{z}_i^{(1)}-\epsilon,0\}}{\partial \widetilde{\mathbf{w}}}$. Then, the following inequality holds:

$$
\begin{aligned}
&\|\frac{\partial G(\mathbf{w}, A_s, B_s)}{\partial \widetilde{\mathbf{w}}}\|^2 = \|\frac{C^{(1)}}{k_1}\sum_{i=1}^{k_1}\iota_i + \frac{C^{(2)}}{k_1}\sum_{i=1}^{k_1}\kappa_i + \frac{C}{k_2}\sum_{i=1}^{k_2}\upsilon_i\|^2 \\
&\leq C^{(1)2}U^{(1)2} + C^{(2)2}U^{(2)2} + C^2\max\{U^{(1)2}, U^{(2)2}\} \\
&+ 2C^{(1)}C^{(2)}U^{(1)}U^{(2)} + 2C^{(1)}CU^{(1)2} + 2C^{(2)}CU^{(2)2}
\end{aligned}
$$

□

**Proof of Theorem 2:** Through calculating the dual of problem (4), it can be concluded that:

$$
\begin{aligned}
&\frac{1}{2}\|\widetilde{\mathbf{w}}\|^2 + \frac{C^{(1)}}{k_1}\sum_{i=1}^{k_1}\xi_i^{(1)} + \frac{C^{(2)}}{k_1}\sum_{i=1}^{k_1}\xi_i^{(2)} + \frac{C}{k_2}\sum_{i=1}^{k_2}\eta_i \\
&\leq -\frac{1}{2}\|\widetilde{\mathbf{w}}\|^2 + \sum_{i=1}^{k_1}\alpha_i^{(1)} + \sum_{i=1}^{k_1}\alpha_i^{(2)} + \epsilon\sum_{i=1}^{k_2}(\alpha_i^{(3)} - \alpha_i^{(4)}), \\
s.t.\quad & 0 \leq \alpha_i^{(1)} \leq \frac{C^{(1)}}{k_1}, \quad 0 \leq \alpha_i^{(2)} \leq \frac{C^{(2)}}{k_1} \\
& \alpha_i^{(3)} \leq 0, \alpha_i^{(4)} \geq 0, \quad 0 \leq \alpha_i^{(4)} - \alpha_i^{(3)} \leq \frac{C}{k_2},
\end{aligned}
$$

where $\alpha_i^{(1)}$, $\alpha_i^{(2)}$, $\alpha_i^{(3)}$, $\alpha_i^{(4)}$ are the dual variables corresponding to the four sets of constraints in problem (3) respectively.

It is clear that,

$$\|\widetilde{\mathbf{w}}\|^2 \leq C^{(1)} + C^{(2)}$$

So, the optimal solution of $\widetilde{\mathbf{w}}$ falls within the ball whose radius is $\sqrt{C^{(1)} + C^{(2)}}$. □

**Proof of Theorem 5:** The Rademacher complexity of the functional space $\mathcal{F}_{C^{(1)}+C^{(2)},D}$ can be upper bounded as follows:

$$
\begin{aligned}
&\hat{\mathcal{R}}_n(\mathcal{F}_{C^{(1)}+C^{(2)},D}) = E_\sigma[\sup_{g\in\mathcal{F}}|\frac{2}{n}\sum_{i=1}^{n}\sigma_i g(\widetilde{\mathbf{B}}_i)|] \\
&=E_\sigma[\sup_{f\in\mathcal{F}}|\frac{2}{n}\sum_{i=1}^{n}\sigma_i\frac{1}{2}(\max_j \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(1)} + \max_j \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij}^{(2)})|] \\
&=E_\sigma[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(\widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij_1^*}^{(1)} + \widetilde{\mathbf{w}}^T\widetilde{\mathbf{B}}_{ij_2^*}^{(2)})|] \\
&\leq\frac{\min\{\sqrt{C^{(1)} + C^{(2)}}, E\}}{n}E_\sigma[|\sum_{i=1}^{n}\sigma_i(\widetilde{\mathbf{B}}_{ij_1^*}^{(1)} + \widetilde{\mathbf{B}}_{ij_2^*}^{(2)})|] \\
&\leq\frac{\min\{\sqrt{C^{(1)} + C^{(2)}}, E\}}{n}\times \\
&\quad\sqrt{\sum_{i=1}^{n}K(\mathbf{B}_{ij_1^*}^{(1)}, \mathbf{B}_{ij_1^*}^{(1)}) + K(\mathbf{B}_{ij_2^*}^{(2)}, \mathbf{B}_{ij_2^*}^{(2)})} \\
&\leq\frac{\min\{\sqrt{C^{(1)} + C^{(2)}}, E\}}{n}\times \\
&\quad(\max_{\rho_{ij}\geq 0,\rho_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_{ij}K(\mathbf{B}_{ij}^{(1)}, \mathbf{B}_{ij}^{(1)})} \\
&\quad+\max_{\rho_{ij}\geq 0,\rho_i^T\mathbf{1}=1}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_{ij}K(\mathbf{B}_{ij}^{(2)}, \mathbf{B}_{ij}^{(2)})})
\end{aligned}
$$

□