2010 IEEE International Conference on Data Mining

A New SVM Approach to Multi-Instance Multi-Label Learning

Nam Nguyen
Cornell University
Department of Computer Science
New York, USA
namn@microsoft.com

***Abstract*—In this paper, we address the problem of multi-instance multi-label learning (MIML) where each example is associated with not only multiple instances but also multiple class labels. In our novel approach, given an MIML example, each instance in the example is only associated with a single label and the label set of the example is the aggregation of all instance labels. Many real-world tasks such as scene classification, text categorization and gene sequence encoding can be properly formalized under our proposed approach. We formulate our MIML problem as a combination of two optimizations: (1) a quadratic programming (QP) that minimizes the empirical risk with L2-norm regularization; and (2) an integer programing (IP) assigning each instance to a single label. We also present an efficient method combining the stochastic gradient decent and alternating optimization approaches to solve our QP and IP optimizations. In our experiments with both an artificially generated data set and real-world applications, i.e. scene classification and text categorization, our proposed method achieves superior performance over existing state-of-the-art MIML methods such as MIMLBOOST, MIMLSVM, M^{3}MIML and MIMLRBF.**

***Keywords*-Multi-Instance Multi-Label, Classification, SVM**

I. Introduction

In recent years, multi-instance multi-label learning (MIML) has attracted significant attention in the machine learning community. MIML is a recently proposed learning framework where each example is associated with a bag of instances as well as a set of labels [1], [2], [3], [4]. There are many real-world applications such as scene classification, text categorization, and gene sequence encoding, which can be formalized under the MIML framework. In scene classification, an image generally partitions into several segments, each can be represented as an instance, while such an image can be labeled into multiple semantic classes simultaneously, such as airplane, ground, building, sky, face, lizard, rock, ... as shown in Figure 1. In text categorization, each document usually comprises of several sections or paragraphs, each can be regarded as an instance, while the document can be assigned to a set of predefined topics such as politics, celebrities, Nobel Prize. In bioinformatics, a gene sequence generally encodes a number of segments, each can be expressed as an instance, while this sequence may be associated with several functional classes, such as metabolism, transcription and protein synthesis.

Figure 1. Scene Classification Examples

In many real-world MIML applications, we observe that given an MIML example, each instance in a bag of instances is mostly associated with a single label and the set of labels of the example is the aggregation of all instance labels. In scene classification, as shown in Figure 1 the object airplane (top left image) is comprised of multiple segments, e.g. two airplane-wing-like segments, an airplane-tail-like segment, and an airplane-body-like segment. These airplane-like segments should be classified into the airplane object class. Similarly, in text categorization each topic in a document is usually described by one or more sections or paragraphs. These sections or paragraphs should be categorized into the topic class that they discuss. In bioinformatics, a functional class in a gene sequence is usually described by several gene segments and these gene segments should be classified into the same functional class as well. Hence, in our approach we make an explicit model assumption that each instance in an MIML examples is associated with exactly a single label.

In previous approaches [1], MIML problem is reduced to its equivalence in the traditional supervised learning, i.e. single-instance single-label learning (SISL) where each example is restricted to have only one instance and only one label. This reduction is done by assigning each instance in a bag of instances to each label in the set of labels. Although

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.109

this transformation from MIML to SISL is feasible, there may be many mislabeled instances. For example, in the airplane image (top left image in Figure 1) the airplane-like segments can be mislabeled as the ground object class and vice versa. Similarly, our proposed approach also reduces MIML problem to SISL problem. By contrast, each instance in the bag of instances is assigned to a single best suitable label in the set of labels in stead of all possible labels in the label sets. In our airplane example, the airplane-like segments should be assigned to a single label: the airplane class. Here, we make an assumption that each instance in an MIML example can be described by at most one semantic class label.

In this paper, we propose a new SVM approach to MIML named SISL-MIML, denoting a novel reduction of MIML problem to the traditional SISL problem. In brief, given an MIML example SISL-MIML seeks the best suitable single label belonging to the set of labels for all instances in the bag of instances simultaneously. Hence, the connections between the instances and labels of an MIML example are explicitly exploited by SISL-MIML. Subsequently, the set of labels of a test example is determined by aggregating all labels of instances in the bag. Each instance is involved in determining the set of labels and the connections between different classes are also addressed in the aggregation phase.

The rest of this paper is organized as follows. Section II reviews the formal definition of MIML and the related work. Section III describes the novel SVM approach to MIML. Section IV reports experimental results on an artificially generated data set as well as the two real-world MIML applications. Finally, Section V summaries and concludes our work.

II. Related Work

In this section, we first reintroduce the formal definition of MIML. In the multi-instance multi-label learning framework, a learning algorithm typically takes a set of labeled training examples $\mathbf{L} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ as input, where $X_i \subseteq \mathcal{X}$ is a bag of instances $\{x_1^i, x_2^i, \ldots, x_{n_i}^i\}$ and $Y_i \subseteq \mathcal{Y}$ is a set of labels $\{y_1^i, y_2^i, \ldots, y_{l_i}^i\}$ associated with X_i. Here n_i is the number of instances in X_i and l_i is the number of labels in Y_i. The goal of MIML is to form a hypothesis $h_{MIML} : 2^{\mathcal{X}} \mapsto 2^{\mathcal{Y}}$ which maps a bag of instances X_i to a set of labels Y_i. The MIML framework can be considered as a generalization from the learning frameworks of multi-instance learning [5], multi-label learning [6], [7] and traditional supervised learning.

Multi-instance learning [5], or multi-instance single-label learning (MISL), was first proposed by Dietterich et al. in their study of predicting the drug molecule activity level. MISL is proposed as a variation of traditional supervised learning framework with incomplete knowledge about labels of training examples. The goal of MISL is to learn a hypothesis $h_{MISL} : 2^{\mathcal{X}} \mapsto \{+1, -1\}$ from a set of MISL training examples $\{(X_i, y_i) \mid 1 \leq i \leq n\}$, where $X_i \subseteq \mathcal{X}$ is a bag of instances $\{x_1^i, x_2^i, \ldots, x_{n_i}^i\}$ and $y_i \in \{+1, -1\}$ is the binary label of the instance X_i. Since the initial work of Dietterich et al. [5], a large number of novel algorithms has been developed in contribution to the development of MISL [8], [9], [10], [11], [12], [13]. In addition, there are many successfully real-world applications of MISL especially in image categorization and retrieval [14], [15], [16], [17], [18]. A more thorough review of multi-instance learning can be found in [19].

Multi-label learning [6], [7], or single-instance multi-label learning (SIML), refers to the classification problem where each example can be assigned to multiple class labels simultaneously. SIML is emerged from the investigation of text categorization problems. The goal of SIML is to learn a mapping $h_{SIML} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ from a set of SIML training examples $\{(x_i, Y_i) \mid 1 \leq i \leq n\}$, where $x_i \in \mathcal{X}$ is a single instance and $Y_i \subseteq \mathcal{Y}$ is a set of labels $\{y_1^i, y_2^i, \ldots, y_{l_i}^i\}$ associated with x_i. SIML has found applications in many different domains, such as natural language processing, computer vision, human computer interaction, bioinformatics, health care, and physiology [20], [21]. There are many existing learning algorithms proposed to exploit the similarity of examples and the correlation among classes [22], [23], [24], [25], [26], [7], [6], [27]. A more thorough review of multi-label learning can be found in [28].

While a large body of work exists on MIML, here we highlight some of the more relevant work. In [1], Zhou and Zhang have formalized the MIML framework and proposed two MIML algorithms named MIMLBOOST and MIMLSVM. The two algorithms transformed MIML into traditional supervised learning (SISL) using MISL and SIML as the connection, respectively. In particular, MIMLBOOST transforms the MIML problem into a multi-instance learning problem (MISL). Each MIML example (X_i, Y_i) is converted into $|\mathcal{Y}|$ number of MISL examples $\{([X_i, y], \mathcal{I}(y \in Y_i) \mid y \in \mathcal{Y}\}$, where $[X_i, y]$ contains n_i instances $\{[x_1^i, y], [x_2^i, y], \ldots, [x_{n_i}^i, y]\}$ formed by concatenating each of X_i's instance with label y, and $\mathcal{I}(y \in Y_i) = \pm 1$ is the corresponding label indicating whether the label y belongs to the set of label Y_i. In order to solve the derived MISL problem, MIMLBOOST employs a specific algorithm named MIBOOSTING [29]. MIBOOSTING reduces the MISL problem into an SISL one under the assumption that each instance in the bag contributes equally and independently to a bag's label. By contrast, MIMLSVM transforms the MIML problem into a multi-label learning problem (SIML). Each MIML example (X_i, Y_i) is converted into an SIML example $(\tau(X_i), Y_i)$, where $\tau(\cdot)$ combines all instances in a bag X_i into a single instance $\tau(X_i)$ using constructive clustering. In order to solve the derived SIML problem, MIMLSVM employs a specific algorithm named MLSVM [20]. This algorithm constructs a binary classifier

for each class label $y \in \mathcal{Y}$ where an instance x_i associated with a label set Y_i is assigned to the positive class $(+1)$ if $y \in Y_i$ otherwise it is assigned to the negative class (-1). Similar to our new proposed algorithm SISL-MIML, these two algorithms [1] transform the MIML problem into its equivalence SISL problem. However, our proposed algorithm is able to exploit the similarity between instances in a bag and the correlation between labels in the label set.

Zhang and Zhou [2] proposed a maximum margin method for the MIML problem named M^{3}MIML. The algorithm assumes a linear model for each class, where the output on one class is set to be the maximum prediction of all the MIML example's instances with respect to corresponding linear model. Subsequently, the outputs on all possible classes are combined to define the margin of the MIML example over the classification system. Similar to our new proposed algorithm SISL-MIML, M^{3}MIML is able to explicitly exploit the connections between the instances and the labels of an MIML examples. In addition, both algorithms utilize the maximum margin learning framework to formulate the optimization problem. However, our proposed algorithm seeks to assign a single best label for each of instances in the example bag simultaneously.

Finally, in [3] the authors developed an innovative neural network style algorithm named MIMLRBF, i.e. Multi-Instance Multi-Label Radial Basis Function. The algorithm is derived from the popular radial basis function method where the first network layer consists of medoids (i.e. bags of instances) formed by performing k-MEDOIDS clustering on MIML examples for each possible class, in which a variant of Hausdorff metric [30] is utilized to measure the distance between bags [31]. Second layer weights of MIMLRBF network are optimized by minimizing a sum-of-squares error function and worked out through singular value decomposition (SVD) [32]. Similar to our new proposed method, connections between instances and labels are directly exploited in the MIMLRBF algorithm. By contrast, our proposed algorithm utilizes a different learning framework, i.e. a maximum margin approach.

III. The New SVM Algorithm of MIML: SISL-MIML

In this section, we discuss how our proposed approach of the MIML problem can be formalized using the maximum margin learning framework. Given a set of MIML labeled training examples $\mathbf{L} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, our proposed algorithm SISL-MIML is also formulated to minimize the regularized empirical risk,

$$\min_w R_{reg}(w) := \lambda\Omega(w) + L(w)$$
$$\text{where } L(w) := \frac{1}{n}\sum_{i=1}^{n} l(X_i, Y_i, w)$$

and where $\Omega(\cdot)$ is a convex and monotonically increasing function which serves as a regularizer with a regularization constant $\lambda > 0$; and $l(X_i, Y_i, w)$ is a nonnegative loss function of an example (X_i, Y_i) measuring the amount of inconsistency between the correct label Y_i and the predicted label arising from using the weight parameter w.

Similar to the Multiclass-SVM proposed by [33], we consider a mapping $\Phi : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{F}$ which projects each instance-label pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ to $\Phi(x, y)$ in a new space $\mathcal{F}$, which is defined as

$$\Phi(x, y) = \begin{bmatrix} x \cdot \mathcal{I}(y = 1) \\ \cdots \\ x \cdot \mathcal{I}(y = i) \\ \cdots \\ x \cdot \mathcal{I}(y = |\mathcal{Y}|) \end{bmatrix},$$

where $\mathcal{I}(\cdot)$ is the indicator function.

In brief, our proposed algorithm, named SISL-MIML, consists of two main components: (1) LABEL-PROPAGATION: given an MIML example we seek the most "suitable" single label for each instance; (2) MARGIN-MAXIMIZATION: given the current labels for all instances, we find the maximum-margin hyperplane that best separates current labeled instances.

In the first step, we make an explicit model assumption that each instance in an MIML examples is associated with exactly a single label. In other words, given an MIML example (X_i, Y_i) our proposed algorithm SISL-MIML seeks the best single label $y_j^i \in Y_i$ for each instance $x_j^i \in X_i$ simultaneously. Based on our model assumption, we also enforce that there should be at least one or more different instances in the bag of instances assigned to each label in the set of labels. Hence, in addition to assigning a single label for each instance, our approach also requires that there are at least $K \geq 1$ number of instances in a bag belonging to a given class label in the set of labels. For example, in Figure 1 there are multiple image segments or patches belonging to each object such as face, airplane, building, lizard, rock, etc. Similarly in text categorization, there should also be multiple or at least one section(s) or paragraph(s) to describe any given topics in a document. Given the set of labels Y_i for a bag of instances X_i, we can view the process of assigning a single label to each instance in the bag as a label propagation procedure which propagates the bag label to each individual instance in the bag. In the margin-based learning framework, given a current weight parameter w and a requirement that there are at least $K \geq 1$ instances belonging to each class label, we can seek the best suitable label assignment for all instances in X_i by solving the following Integer Programming (IP), called LABEL-PROPAGATION:

OPTIMIZATION I: LABEL-PROPAGATION

$$\min_{\overline{y}^i_1,\ldots,\overline{y}^i_{n_i}} \sum_{x^i_j \in X_i} \left| \max_{\hat{y}^i_j \in \mathcal{Y}} \left[w^T \Phi(x^i_j, \hat{y}^i_j) \right] - w^T \Phi(x^i_j, \overline{y}^i_j) \right| \quad (1)$$

subject to:

$$\sum_{j=1}^{n_i} \mathcal{I}(\overline{y}^i_j = y^i) \geq K, \text{ for all } y^i \in Y_i,$$

$$\overline{y}^i_j \in Y_i, \text{ for all } 1 \leq j \leq n_i,$$

where $\mathcal{I}(\cdot)$ is the indicator function; $\sum_{j=1}^{n_i} \mathcal{I}(\hat{y}^i_j = y^i)$ is the number of instances assigned to the class label $y^i \in Y_i$; $w^T \Phi(x^i_j, \overline{y}^i_j)$ is the score associated with the most "suitable" label $\overline{y}^i_j$; and $\max_{\hat{y}^i_j \in \mathcal{Y}} \left[w \Phi(x^i_j, \hat{y}^i_j) \right]$ is the maximum score value associated with any labels $\hat{y}^i_j \in \mathcal{Y}$. In Equation 1, the objective of the given LABEL-PROPAGATION formulation seeks the labels belong to the set of label Y_i whose associated scores are as large as the maximum score. Furthermore, the LABEL-PROPAGATION integer programing also exploits the connection between instances and labels of a given MIML example by enforcing the constraint that there are at least $K \geq 1$ instances belonging to each class label and ensuring all instance labels belong to the set of correct labels. Consequently, the solution of the LABEL-PROPAGATION optimization in Equation 1, $\overline{Y}_i = \{\overline{y}^i_1, \ldots, \overline{y}^i_{n_i}\}$, can be viewed as the approximated true labels of all instances in a bag of instances X_i given the current weight parameter w, which can be used to compute the loss function $l(X_i, Y_i, w)$ in the next step. In other words, our first step can be viewed as a relaxation of an explicit division of all instances into different groups which is assigned to different labels in the label set.

In the second step, the SISL-MIML algorithm can be obtained by considering the situation where we use the L2-norm regularization,

$$\Omega(w) = \frac{1}{2} \|w\|^2,$$

and the loss function $l(X_i, Y_i, w)$ is set to the average hinge loss of all instances in the bag of instances X_i,

$$\frac{1}{n_i} \sum_{x^i_j \in X_i} \max(0, 1 - [w^T \Phi(x^i_j, \overline{y}^i_j) - \max_{\hat{y}^i_j \neq \overline{y}^i_j} w^T \Phi(x^i_j, \hat{y}^i_j)]),$$

where $\overline{Y}_i = \{\overline{y}^i_1, \ldots, \overline{y}^i_{n_i}\}$ is the approximate true labels of all instances in the bag of instances (i.e. the solution of integer programming 1). Specifically, the SISL-MIML learns a weight vector w and slack variables ξ via the following quadratic optimization:

OPTIMIZATION II: MARGIN-MAXIMIZATION

$$\min_{w, \xi \geq 0} : \frac{\lambda}{2} \|w\|^2 + \frac{1}{\sum_{i=1}^n n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \xi^i_j \quad (2)$$

subject to:

$\forall (X_i, Y_i) \in \mathbf{L}$ and $\forall x^i_j \in X_i$:

$$w^T \Phi(x^i_j, \overline{y}^i_j) - \max_{\hat{y}^i_j \neq \overline{y}^i_j} w^T \Phi(x^i_j, \hat{y}^i_j) \geq 1 - \xi^i_j,$$

where $\overline{Y}_i = \{\overline{y}^i_1, \ldots, \overline{y}^i_{n_i}\}$ is the solution of the LABEL-PROPAGATION in equation 1.

In the testing phase, after we have learned the weight parameter w and slack variables ξ, the classification of a test MIML example X_t is done by

$$h_{MIML}(X_t) = \{y_t \in \mathcal{Y} \mid \sum_{\hat{y}^t_j \in \overline{Y}_i} \mathcal{I}(\hat{y}^t_j = y_t) \geq K\}.$$

Both of our optimization formulations, LABEL-PROPAGATION and MARGIN-MAXIMIZATION, are depended on the solution of each other. In the LABEL-PROPAGATION, given the weight parameter w, the given integer programming is a convex optimization. Similarly, in the MARGIN-MAXIMIZATION formulation, given the assigned labels of instances for all examples, $\overline{Y}_i$, $\forall\ 1 \leq i \leq n$, the given quadratic programming is also a convex optimization. Hence, in order to solve the proposed quadratic programming, we employ both the alternating optimization procedure [34] and the stochastic gradient descent approach which has shown to be very efficient and does not require transforming to the dual formulation [35], [36]. Here, we embed the integer programming problem in the stochastic gradient decent procedure to solve the quadratic programming problem. Similar to [36], we restrict the search space to the sphere of radius $1/\sqrt{\lambda}$. The algorithm alternates between a gradient descent step which also include solving the LABEL-PROPAGATION integer programming, and a projection step until reduction of the regularized risk objective function is less than a pre-specified tolerance, ϵ. In each iteration, for each MIML example (X_i, Y_i) the algorithm first solves the LABEL-PROPAGATION integer programing to obtain the approximated true labels $\overline{Y}_i = \{\overline{y}^i_1, \ldots, \overline{y}^i_{n_i}\}$. Secondly, the algorithm computes instances that violate the constraints in the MARGIN-MAXIMIZATION optimization problem. Then the weight parameter w is updated according to the violated instances found in the previous step. In the projection step, the weight parameter w is projected to the sphere of radius $1/\sqrt{\lambda}$. The details of the SISL-MIML algorithm are given in Algorithm 1.

In order to use the *kernel trick*, as pointed out in [36], we set $w_1 = 0$ then w_t can be written as

$$w_t = \sum_{(x,y)} \varphi_{xy} \Phi(x, y),$$

where φ_{xy} is a scalar associated with a vector $\Phi(x, y)$. Hence, we can incorporate the usage of kernel when com-

Algorithm 1 : MIML Algorithm: SISL-MIML

Input: **L** - the fully labeled data
λ - the regularization constant
K - instances belonging to each class label
ϵ - a tolerance for stopping condition

Initialize w_1 such that $\|w_1\| \leq 1/\sqrt{\lambda}$ and set $t = 1$
repeat
for each $(X_i, Y_i) \in \mathbf{L}$

Given the weight w_t and K, let $\overline{Y}_i = \{\overline{y}_1^i, \ldots, \overline{y}_{n_i}^i\}$ be the solution of the LABEL-PROPAGATION integer programming

Set $\eta_t = \dfrac{1}{\lambda t}$ and $w_{t+1/2} = (1 - \eta_t \lambda) w_t$

// Find violated constraints in QP and update w
for each $x_j^i \in X_i$

if $\left(w_t^T \Phi(x_j^i, \overline{y}_j^i) - \max\limits_{\hat{y}_j^i \neq \overline{y}_j^i} w_t^T \Phi(x_j^i, \hat{y}_j^i) < 1 \right)$

Set $w_{t+1/2} = w_{t+1/2} + \dfrac{\eta_t}{\sum_{i=1}^n n_i} \left[\Phi(x_j^i, \overline{y}_j^i) - \Phi(x_i, \hat{y}_j^i) \right]$

where $\hat{y}_j^i = \arg\max\limits_{\hat{y}_j^i \neq \overline{y}_j^i} w_t^T \Phi(x_j^i, \hat{y}_j^i)$

end if
end for

// Project w to the ball of radius $1/\sqrt{\lambda}$
Set $w_{t+1} = \min\left\{1, \dfrac{1/\sqrt{\lambda}}{\|w_{t+1/2}\|}\right\} w_{t+1/2}$
Set $t = t + 1$
end for
until $(R_{reg}(w_{t-1}) - R_{reg}(w_t)) < \epsilon$

Output: w_t

puting inner product operations, i.e.:

$$\begin{aligned} \langle w, \Phi(x', y') \rangle &= \sum_{(x,y)} \varphi_{x,y} \mathbf{K}(x, y, x', y') \\ \|w\|^2 &= \sum_{(x,y)} \sum_{(x',y')} \varphi_{x,y} \varphi_{x',y'} \mathbf{K}(x, y, x', y') \end{aligned}$$

In our experiment, we use the radial basis function (RBF) as a kernel,

$$\mathbf{K}(x, y, x', y') = \exp\left(-\frac{\|\Phi(x, y) - \Phi(x', y')\|^2}{2\sigma^2} \right),$$

where the radius σ determines the smoothness of the decision boundary.

IV. Experiments

In this section, we compare performance of our proposed method SISL-MIML with other recently developed MIML algorithms: MIMLRBF [3], M^{3}MIML [2], and MIMLSVM, MIMLBOOST [1]. For fair comparison, the RBF kernel is used for all MIML algorithms with the radius $\sigma = 0.2$. Specifically, the MIMLRBF algorithm involves two different parameters: the fraction parameter α and the scaling factor μ which are determined by two-fold cross validation on the training examples where $\alpha \in \{10\%, 20\%, 30\%\}$ and $\mu \in \{0.5, 0.6, 0.7\}$. In addition, the M^{3}MIML algorithm requires two different parameters: the cost parameter C which is set to the best values in the range $\{10^i \mid -4 \leq i \leq 4\}$ by two-fold cross validation using the training examples and γ which is set to the default value of 1. Furthermore, the parameters for the two algorithms MIMLSVM and MIMLBOOST is set according the best values as reported in [1]. Finally, our proposed algorithm SISL-MIML also associates two different parameters: (1) the regularization constant, λ; and (2) the number of instances required to belong to a single label class, K. Both of these parameters are determined by two-fold cross validation on the training examples where $\lambda \in \{10^i \mid -4 \leq i \leq 4\}$ and $K \in \{1, 2, 3, 4\}$.

To evaluate the performance of different MIML algorithms we look at a set of five standard multi-label performance measures: Hamming Loss, One Error, Coverage, Ranking Loss, and Average Precision. For Average Precision, the bigger the value the better the performance. While for the other four measures, the smaller the value the better the performance. A more detailed description of these five performance measures can be found in [7], [25].

We evaluate performance of different MIML algorithms on an artificially generated data set and two real-word MIML applications. The artificially generated data set is originated from the USPS data in the UCI repository [37] which contains $9K$ SISL examples of 256 features belonging to 10 possible class labels. The generated data contains $4K$ MIML examples for training and $10K$ MIML examples for testing. We generated multiple versions of the MIML data set in which we vary the number of labels per example, i.e. the size of the set of labels for each MIML example, and instances per label, i.e. the number of instances belongs to each label in the set of labels for each MIML example. Hence, the number of instances per example is the product of the number of instances per label and labels per example. In our experiment, the number of labels per example and instances per label takes values from 2-to-5 and 2-to-4, respectively. Given the number of labels per example and instances per label, i.e. l_1 and l_2, an MIML example is generated as follow: (1) first we randomly generate l_1 different labels from the set of all possible labels $\mathcal{Y}$ based on the uniform distribution and (2) for each generated label we make an

Table I
PROPERTIES OF THE TWO REAL-WORLD MIML APPLICATIONS: IMAGE AND TEXT DATA SETS.

DATA SET	NUMBER OF EXAMPLES	NUMBER OF CLASSES	NUMBER OF FEATURES	INSTANCES PER EXAMPLE			LABELS PER EXAMPLE		
				MIN	MAX	MEAN±STD.	$k=1$	$k=2$	$k\geq 3$
IMAGE	2000	5	15	9	9	9.00±0.00	1543	442	15
TEXT	2000	7	243	2	26	3.56±2.71	1701	290	9

uniformed random selection of l_2 instances belonging to the given class label to form the MIML example. In addition, the data sets for the two real-world applications are (1) the image data set collected from the COREL image collection and the Internet; and (2) the text data set collected from the widely studied Reuters-21578 collection [38]. In Table I, we present a brief characteristic description of the two real-world data sets. A more detail description of the data sets are found in [3], [2]. For these two data sets, performance of different MIML algorithms is reported based on ten-fold cross validation.

In Figure 2 (top row), we plot the overall average performance of MIML algorithms among different values of the number of labels per example and instances per label on the USPS artificially generated data set. Across different performance metrics, our proposed algorithm SISL-MIML consistently produces superior performance against other MIML methods. Especially there is a significant margin of improvement in Coverage, Ranking Loss and Average Precision performance measures. The improvement in performance demonstrates that our proposed algorithm SISL-MIML is able to take advantage of our explicit model assumption since the artificially generated data is constructed based on this assumption.

Furthermore, in Figure 2 (bottom two rows), we plot the average performance (mean $\pm$ standard error) of MIML algorithms on both the image and text data sets. The new SISL-MIML algorithm consistently yields results equal to or better than MIMLRBF and M^{3}MIML. Similar to [3], [2], we also observe the significant improvement in performance of both MIMLRBF and M^{3}MIML over MIMLSVM and MIMLBOOST. In addition, we also observe that SISL-MIML is able to outperform both MIMLRBF and M^{3}MIML on the image data set. This behavior is confirmed our model assumption that each segment in an image is associated with only a single class label.

Moreover, we also investigate the behavior of different MIML algorithms as we vary the number of labels per example and instances per label of the artificially generated data set. As shown in Figure 3 and 4, we plot the average performance of MIML algorithms versus the number of labels per example and instances per label, respectively. As the number of labels per example increases, performance on Hamming Loss and Coverage gets worse, while performance on One Error and Average Precision improves. These conflicting trends demonstrate that the size of the set of labels affects different performance measures differently. For example, since the Hamming Loss measures the intersection between the true set of labels and the predicted set of labels, the measure would reflect the difficulty of the MIML problem when the set of labels increases in size. Since the One Error only pays attention to the instance with the highest confidence, the measure reflects the easiness of the MIML problem when the set of labels increases in size. By contrast, as the number of instances per label varies, the performance measures of all MIML algorithms do not seem to be affected. The effect can be explained by the fact that the number of instances belonging to a same class label in each MIML example may not contribute in the process of determining the set of predicted labels. Furthermore, to determine whether a class label belongs to the set of predicted labels for an MIML example, the deciding factor is that there exists an instance of that class label.

V. SUMMARY

In this paper, we introduce SISL-MIML, a novel SVM method for the MIML problem. Our proposed algorithm reduces the MIML problem to the traditional SISL problem. Specifically, given an MIML example, SISL-MIML seeks the best suitable single label belonging to the set of labels for all instances in the bag of instances simultaneously. Hence, the connections between the instances and labels of an MIML example are explicitly exploited by SISL-MIML. Experiments using both the artificially generated data and two real-world applications shows that SISL-MIML are able to produce superior performance in comparison to other MIML algorithms. In addition, using the artificial generated data we also investigate the behavior of different MIML algorithms when we vary the number of labels per examples and instances per label.

REFERENCES

[1] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *The Neural Information Processing Systems (NIPS)*, 2006, pp. 1609–1616.

[2] M.-L. Zhang and Z.-H. Zhou, "M3MIML: A maximum margin method for multi-instance multi-label learning," in *ICDM*, 2008, pp. 688–697.

[3] M.-L. Zhang and Z.-J. Wang, "MIMLRBF: RBF neural networks for multi-instance multi-label learning," *Neurocomputing*, vol. 72, no. 16-18, pp. 3951–3956, 2009.

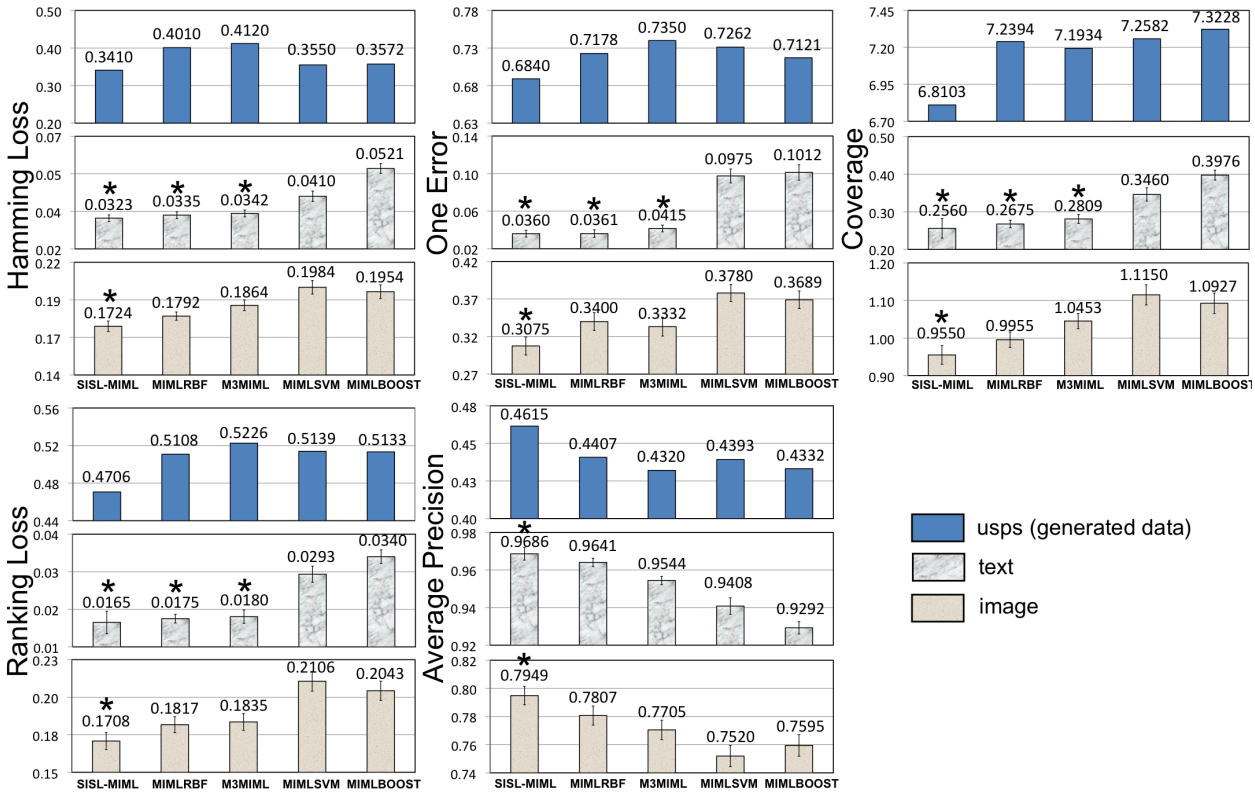

Figure 2. Overall average performance of different MIML algorithms: SISL-MIML, MIMLRBF, M^{3}MIML, MIMLSVM, and MIMLBOOST, across different evaluation metrics: Hamming Loss, One Error, Coverage, Ranking Loss, and Average Precision. The error bar indicates standard error and the star(*) indicates statistically significant improvement at the $\alpha = 0.05$ level according to a two-sided t-test.

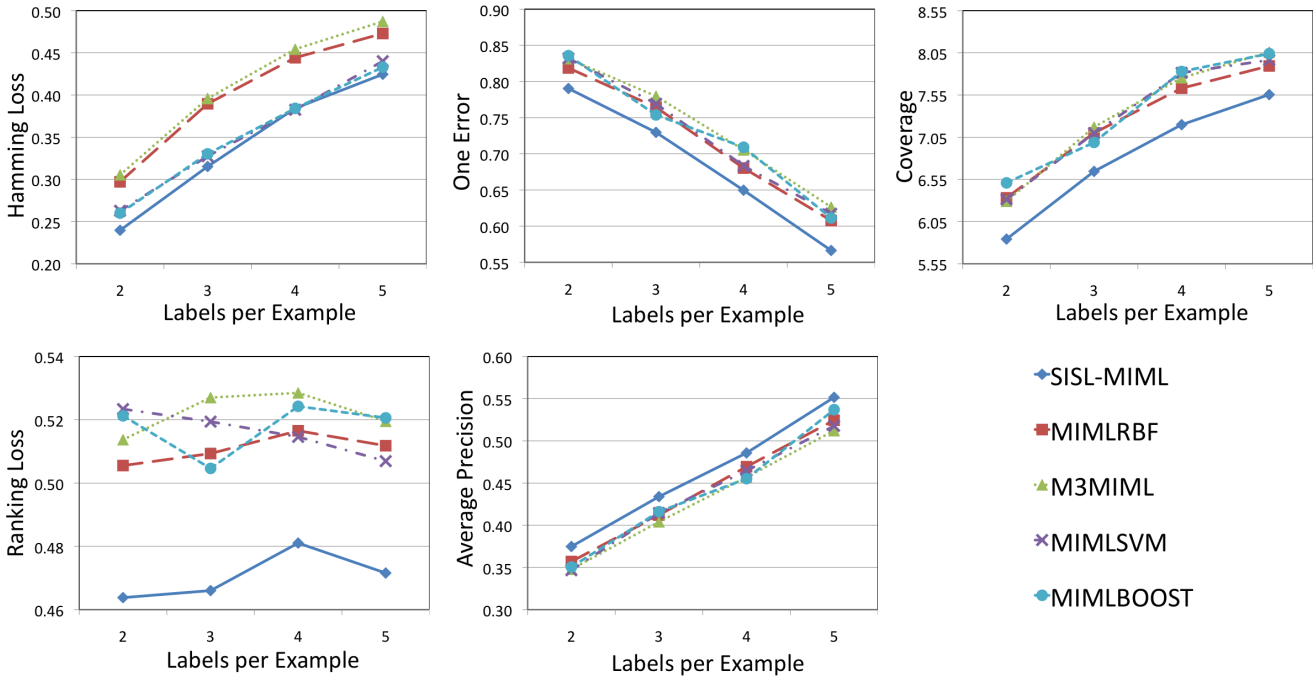

Figure 3. Average performance of different MIML algorithms when the number of labels per example is varied.

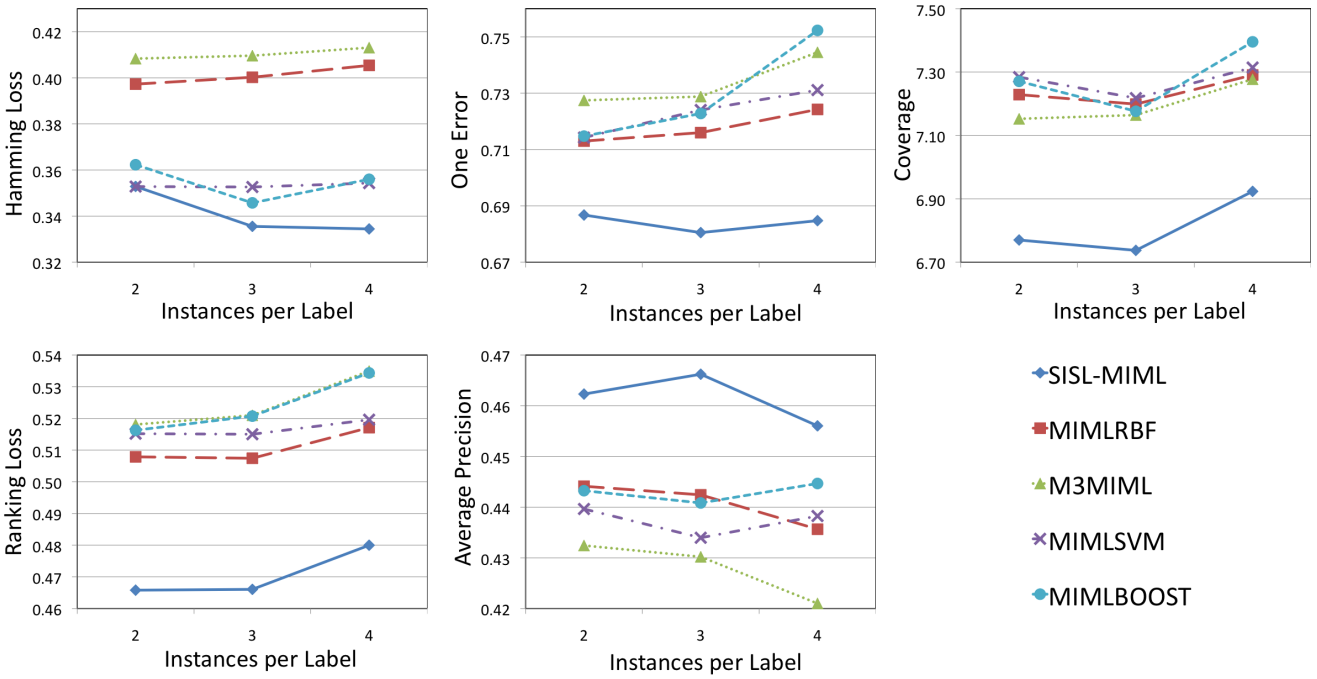

Figure 4. Average performance of different MIML algorithms when the number of instances per label is varied.

[4] Z. J. Zha, X. S. Hua, T. Mei, J. D. Wang, G. J. Qi, and Z. F. Wang, "Joint multi-label multi-instance learning for image classification," in *CVPR*, 2008, pp. 1–8.

[5] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[6] A. McCallum, "Multi-label text cassification by EM," *AAAI'99 Workshop on Text Learning.*, 1999.

[7] R. E. Schapire and Y. Singer, "BOOSTEXTER: a boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[8] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *The Neural Information Processing Systems (NIPS)*, 2002, pp. 561–568.

[9] Y. Chevaleyre and J.-D. Zucker, "Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. application to the mutagenesis problem," *Lecture Notes in Computer Science*, vol. 2056, pp. 204–214, 2001.

[10] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *The Neural Information Processing Systems (NIPS)*, 1997.

[11] S. Andrews and T. Hofmann, "Multiple-instance learning via disjunctive programming boosting," in *The Neural Information Processing Systems*, 2003.

[12] S. Ray and D. Page, "Multiple instance regression," in *Proceedings of The International Conference on Machine Learning (ICML)*, 2001, pp. 425–432.

[13] J. Wang, et Jean-Daniel Zucker, and J. daniel Zucker, "Solving the multiple-instance problem: A lazy learning approach," in *Proceedings of The International Conference on Machine Learning (ICML)*, 2000, pp. 1119–1125.

[14] Y. X. Chen, J. B. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[15] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *JMLR*, vol. 5, pp. 913–939, 2004.

[16] Q. Zhang, S. A. Goldman, W. Yu, and J. E. Fritts, "Content-based image retrieval using multiple-instance learning," in *Proceedings of The International Conference on Machine Learning (ICML)*, 2002, pp. 682–689.

[17] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *Image Understanding Workshop*, 1998, pp. 1031–1036.

[18] Xu, Jinxi, Croft, and W. Bruce, "Query expansion using local and global document analysis," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. Experimental Studies, 1996, pp. 4–11.

[19] J. Yang, "Review of multi-instance learning and its application," 2001. [Online]. Available: http://www.cs.cmu.edu/ juny/MILL/review.htm

[20] M. R. Boutell, J. B. Luo, X. P. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[21] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A MFoM learning approach to robust multiclass multi-label text categorization," in *Proceedings of The International Conference on Machine Learning (ICML)*, 2004.

[22] K. Brinker, J. Fürnkranz, and E. Hüllermeier, "A unified model for multilabel classification and ranking," in *UCAI*, vol. 141, 2006, pp. 489–493.

[23] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *The Neural Information Processing Systems (NIPS)*, 2001, pp. 681–687.

[24] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda, "Maximal margin labeling for multi-topic text categorization," in *The Neural Information Processing Systems (NIPS)*, 2004.

[25] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[26] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 274–281.

[27] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," in *The Neural Information Processing Systems (NIPS)*, 2002, pp. 721–728.

[28] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *IJDWM*, vol. 3, no. 3, pp. 1–13, 2007.

[29] X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," pp. 272–281, 2004.

[30] G. A. Edgar, *Measure, topology, and fractal geometry*. New York: Springer, 1990.

[31] M.-L. Zhang and Z.-H. Zhou, "Multi-instance clustering with applications to multi-instance prediction," *Application Intell*, vol. 31, no. 1, pp. 47–68, 2009.

[32] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 1986.

[33] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[34] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *Neural, Parallel & Scientific Computations*, vol. 11, no. 4, pp. 351–368, 2003.

[35] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *Proceedings of The International Conference on Machine Learning (ICML)*, 2006, pp. 969–976.

[36] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for svm," in *Proceedings of the 24th International Conference on Machine Learning*. New York, NY, USA: ACM, 2007, pp. 807–814.

[37] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[38] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.