Towards Discovering What Patterns Trigger What Labels *

Yu-Feng Li Ju-Hua Hu Yuan Jiang Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210046, China

Abstract

In many real applications, especially those involving data objects with complicated semantics, it is generally desirable to discover the relation between patterns in the input space and labels corresponding to different semantics in the output space. This task becomes feasible with MIML (Multi-Instance Multi-Label learning), a recently developed learning framework, where each data object is represented by multiple instances and is allowed to be associated with multiple labels simultaneously. In this paper, we propose KISAR, an MIML algorithm that is able to discover what instances trigger what labels. By considering the fact that highly relevant labels usually share some patterns, we develop a convex optimization formulation and provide an alternating optimization solution. Experiments show that KISAR is able to discover reasonable relations between input patterns and output labels, and achieves performances that are highly competitive with many state-of-the-art MIML algorithms.

Introduction

In many real applications, especially in applications involving data objects with complicated semantics, besides achieving a high performance for prediction, it is usually desirable to discover the relation between the input patterns and output labels because it is helpful to understand the semantic formations and may lead to advanced technical designs. For example, in text applications, the discovery of relations between words and topics may help improve the performance of question answering systems (Voorhees 2003); in image applications, the discovery of relations between segmented regions and tags may help improve the ability of recognizing human movements (Gavrila 1999); in audio applications, the discovery of relations between voices and persons may help expand applicability of hands-free computing (Omologo *et al.* 1998); furthermore, in video applications, the discovery of relations between flames and annotations may help design more effective tracking systems (Lipton *et al.* 1998).

There lacks a general learning framework for this purpose for a long time. For instance, in traditional supervised learning, each pattern is related to only one class label which generally corresponds to a singe semantic; in multi-label learning, although different semantics can be encoded by multiple class labels, they are always triggered by the identical single input pattern. Recently, Zhou *et al.* (2012; 2006) proposed MIML (Multi-Instance Multi-Label learning), a new learning framework where each data object is represented by multiple instances and is allowed to be associated with multiple labels simultaneously. As illustrated in Figure 1, MIML explicitly considers patterns in the input space and labels corresponding to different semantics in the output space, and thus, it enables to discover the relation between input patterns and output labels.

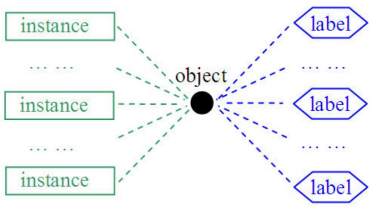

Figure 1: Illustration of the MIML framework (Zhou *et al.* 2012; Zhou and Zhang 2006).

During the past few years, many MIML algorithms have been developed (Zhou and Zhang 2006; Zha *et al.* 2008; Zhang and Zhou 2009; Jin *et al.* 2009; Yang *et al.* 2009; Nguyen 2010; Luo and Orabona 2010; Zhou *et al.* 2012). To name a few, Zhou and Zhang (2006) proposed MIMLSVM by degenerating MIML to single-instance multi-label learning and MIMLBOOST by degenerating MIML to multi-instance single-label learning. Zha *et al.* (2008) proposed an MIML algorithm based on hidden conditional random field. Yang *et al.* (2009) formulated MIML as a probabilistic generative model. MIML nearest neighbor and neural network algorithms have been presented in (Zhang 2010; Zhang and Wang 2009), and MIML metric learning has been studied in (Jin *et al.* 2009). It is noteworthy that most previous studies focused on improving generalization, whereas few considers the potential of MIML in discovering the pattern-label relations although the possibility has been pointed out by (Zhou *et al.* 2012).

In this paper, we propose KISAR (Key Instances Sharing Among Related labels). We consider that a certain label is

*This research was supported by NSFC (61073097, 60975043), 973 Program (2010CB327903) and Baidu fund.

associated to a *bag* of instances because there are some instances, rather than all the instances in the bag, triggering this label; we refer to these instances as *key instances* for the label. We also consider that highly relevant labels generally share some key instances. Thus, a two-stage approach is developed. We first study a mapping from a bag of instances to a feature vector, where each feature value measures the degree of the bag being associated to a group of similar instances. Sparse predictors are then employed for labels to select key groups of instances and consequently key instances. The task is formulated as a convex optimization problem and an alternating optimization solution is provided. Experimental results show that KISAR is able to discover reasonable relations between input patterns and output labels, as well as achieve predictive performances highly competitive with many state-of-the-art MIML algorithms.

In the following we start by presenting KISAR, and then report experimental results before the conclusion.

The KISAR Approach

In MIML, each example is a bag of instances and associated with multiple class labels. Formally, we are given a set of training examples $\mathcal{D} = \{B_i, Y_i\}_{i=1}^n$, where $B_i = \{\mathbf{x}_{i,j}\}_{j=1}^{m_i}$ is a bag containing m_i instances, $Y_i = [y_{i,1}, \ldots, y_{i,T}] \in \{\pm 1\}^T$ is the label vector of bag B_i, $y_{i,t} = +1$ indicates that the label t is proper for the bag B_i, n is the number of training examples and T is the total number of class labels. The goal is to predict all proper labels for any unseen bag, noticing that the number of proper labels is unknown. Mathematically, let $\mathcal{X}$ and $\mathcal{Y}$ be the space of $\mathbf{x}$ and y, then the goal is to learn a decision function $F : 2^{\mathcal{X}} \to 2^{\mathcal{Y}}$ based on $\mathcal{D}$.

To discover the instance-label relations, we consider a mapping from a bag of instances to a feature vector :

$$\phi(B) = [sim(B, c_1), \ldots, sim(B, c_d)], \qquad (1)$$

where $c_1, \ldots, c_d$ are d prototypes of all the instances, and sim is a similarity function. The larger the value of $sim(B, c_p)$ ($\forall p \in \{1, \cdots, d\}$), the more similar the bags B and c_p. With this mapping, a bag is represented by a traditional single feature vector and thus, classical supervised learning algorithms are readily applicable. In particular, when supervised learning algorithms are used to select the key prototypes, they may also be capable of identifying the key instances since each prototype generally corresponds to a group of similar instances. In other words, with such a mapping, the task of identifying key instances is alleviated by a simpler task of identifying key prototypes.

To instantiate the mapping, in this paper we set the prototypes as cluster centers via k-means and the similarity function as Gaussian distance which is naturally normalized, *i.e.*, $sim(B, c) = \min_{\mathbf{x} \in B} \exp(-\frac{\|\mathbf{x}-c\|_2^2}{\delta})$ where δ is fixed to be the average distance between the instances in one cluster. A similar process has been used in (Zhou and Zhang 2007; Zhou *et al.* 2012). Notice that other implementations are also possible. For example, in MILES (Chen *et al.* 2006), all the instances are viewed as prototypes and then the similarity is calculated via *most-likely-cause estimator* (Maron and Lozano-Pérez 1998); in DD-SVM (Chen and Wang 2004), non-convex Diversity-Density (DD) algorithm (Maron and Lozano-Pérez 1998) is executed for multiple times, the multiple local maximal solutions are recorded as the prototypes, and finally the minimal Euclidean distance is used for the similarity function.

A Convex Formulation

One direct approach to identify the key instances is to treat the labels in an independent manner. Such an approach, however, ignores the fact that highly relevant labels usually share common key instances, and thus leading to suboptimal performances. By contrast, in the following we explicitly take label correlations into account.

Let $G \in [0, 1]^{T \times T}$ denote a label relation matrix, *i.e.*, $G_{t,\hat{t}} = 1$ means that labels t and $\hat{t}$ are related and 0 otherwise[1]. The direct approach can be viewed as a special case by initializing G as an identity matrix. Consider a decision function $F = [f_1, \ldots, f_T]$, where each f_t corresponds to a single label. Without loss of generality, suppose that each f_t is a linear model, *i.e.*, $f_t = \mathbf{w}_t' \phi(B)$ where $\mathbf{w}_t = [w_{t,1}, \ldots, w_{t,d}] \in \mathcal{R}^{d \times 1}$ is the linear predictor and $\mathbf{w}_t'$ denotes the transpose of $\mathbf{w}_t$. Then our goal is to find $\mathbf{w}_t$'s such that the following functional is minimized,

$$\min_{\mathbf{W}=[\mathbf{w}_1,\ldots,\mathbf{w}_T]} \gamma \sum_{i=1}^{n} \sum_{t=1}^{T} \ell(y_{i,t}, \mathbf{w}_t' \phi(B_i)) + \Omega(\mathbf{W}, G), \quad (2)$$

where ℓ is a convex loss function, *e.g.*, the hinge loss $\ell(y_{i,t}, \mathbf{w}_t' \phi(B_i)) = \max\{0, 1 - y_{i,t} \mathbf{w}_t' \phi(B_i)\}$ used in this paper, $\Omega(\mathbf{W}, G)$ is a regularization term that characterizes the consistency between the predictors $\mathbf{W}$ and label relation matrix G, and γ is a parameter trading-off the empirical loss and model regularization.

Let $\mathbf{w}^{(t,\hat{t})}$ denote $[\mathbf{w}_t, \mathbf{w}_{\hat{t}}]$ for any related-label pair $(t, \hat{t})$. The assumption of common key instances implies that many rows of $\mathbf{w}^{(t,\hat{t})}$ are identically equal to zero, *i.e.*, these corresponding prototypes are not useful for both labels t and $\hat{t}$. Notice that the number of non-zero rows in matrix $\mathbf{w}^{(t,\hat{t})}$ is computed by $\|\mathbf{w}^{(t,\hat{t})}\|_{2,0}$ where $\|\mathbf{w}^{t,\hat{t}}\|_{2,0} = |\sqrt{w_{t,1}^2 + w_{\hat{t},1}^2}, \ldots, \sqrt{w_{t,d}^2 + w_{\hat{t},d}^2}|_0$. This motivates the definition of $\Omega(\mathbf{W}, G)$ as $\sum_{1 \le t, \hat{t} \le T} G_{t,\hat{t}} \|\mathbf{w}^{(t,\hat{t})}\|_{2,0}^2$, and therefore Eq. 2 can be rewritten as

$$\min_{\mathbf{W}} \gamma \sum_{i=1}^{n} \sum_{t=1}^{T} \ell(y_{i,t}, \mathbf{w}_t' \phi(B_i)) + \sum_{1 \le t, \hat{t} \le T} G_{t,\hat{t}} \|\mathbf{w}^{(t,\hat{t})}\|_{2,0}^2. \quad (3)$$

Here the entries of G can be viewed as multiple regularization parameters controlling the sparsity of each label pair; the larger the $G_{t,\hat{t}}$ value, the sparser the $\mathbf{w}^{(t,\hat{t})}$.

The optimization problem in Eq.3, however, is non-continuous and non-convex. With the popular and widely-accepted convex relaxation approach, we can replace the non-continuous and non-convex term $\|\mathbf{w}^{(t,\hat{t})}\|_{2,0}$ with a

[1]Here G can be extended to a weighted label relation matrix.

tight convex function, *e.g.*, $L_{2,1}$ norm $\|\mathbf{w}^{(t,\hat{t})}\|_{2,1}$, *i.e.*,

$$\min_{\mathbf{W}} \gamma \sum_{i=1}^{n}\sum_{t=1}^{T} \ell(y_{i,t}, \mathbf{w}_t'\phi(B_i)) + \sum_{1\le t,\hat{t}\le T} G_{t,\hat{t}} \|\mathbf{w}^{(t,\hat{t})}\|_{2,1}^2. \quad (4)$$

In (Donoho 2006) and references therein, it has been indicated that L_1 norm is a tight relaxation to L_0 norm, *i.e.*, under some mild conditions of the loss function, the solution of L_1 relaxation could be equivalent to that of L_0 problem. It is noteworthy that the form in Eq.4 can be viewed as an extension to the *common feature learning* of multi-task learning (Argyriou *et al.* 2008) but with a different purpose. Specifically, different to (Argyriou *et al.* 2008) which assumes that *all* the tasks share the same common features, here we consider a more flexible way that the common features shared by different related-labels can be different; moreover, they were in the setting of multi-task learning whereas we are working on discovering pattern-label relation in MIML.

Alternating Optimization Solution

It appears that state-of-the-art convex optimization techniques can be readily applied to solve the convex problem in Eq.4. Due to the non-smoothness of the $L_{2,1}$ term, however, Eq.4 is challenging to solve in an efficient manner. To tackle this difficulty, two efficient approaches have been developed. One is alternating optimization algorithm (Argyriou *et al.* 2008) which first rewrites the $L_{2,1}$ term as a jointly-convex function of $\mathbf{W}$ with additional variables $\boldsymbol{\lambda}$, and then optimizes $\mathbf{W}$ and $\boldsymbol{\lambda}$ iteratively until convergence. Another approach is based on accelerated gradient algorithms (Liu *et al.* 2009; Ji and Ye 2009) using optimal first-order method (Nesterov and Nesterov 2004). It can be shown that accelerated gradient algorithms achieve the optimal convergence rate, *i.e.*, $O(1/s^2)$ where s is the number of iterations; moreover, the computational cost of each step is only related to an Educlidean projection that is also efficient. All these nice properties, however, only hold for functions containing a single $L_{2,1}$ term, obviously not the case in Eq.4, where there are usually multiple $L_{2,1}$ terms in the objective. Alternatively, in the following we show that the alternating optimization algorithm does not suffer from this issue and can be adapted to solve Eq.4. We first introduce a theorem.

Theorem 1 *(Argyriou* et al. *2008) Use the convention that* $\frac{x}{0} = 0$ *if* $x = 0$ *and* ∞ *otherwise. Then we have,*

$$\|\mathbf{w}^{(t,\hat{t})}\|_{2,1}^2 = \min_{\boldsymbol{\lambda}^{(t,\hat{t})}\in\mathcal{M}} \left(\langle \mathbf{w}_t, D_{t,\hat{t}}^{+}\mathbf{w}_t\rangle + \langle \mathbf{w}_{\hat{t}}, D_{t,\hat{t}}^{+}\mathbf{w}_{\hat{t}}\rangle\right)$$

where $\mathcal{M} = \{\boldsymbol{\lambda} | \lambda_i \ge 0, \sum_{i=1}^{d}\lambda_i \le 1\}$, $D_{t,\hat{t}} = Diag(\boldsymbol{\lambda}^{(t,\hat{t})})$, $D_{t,\hat{t}}^{+}$ *denotes the pseudoinverse of* $D_{t,\hat{t}}$. *Here* $\langle , \rangle$ *denotes the inner product.*

According to Theorem 1, Eq.4 can be rewritten as:

$$\min_{\mathbf{W}} \min_{\{\boldsymbol{\lambda}^{(t,\hat{t})}\}_{t,\hat{t}=1}^{T}} \gamma \sum_{i=1}^{n}\sum_{t=1}^{T} \ell(y_{i,t}, \mathbf{w}_t'\phi(B_i)) + \sum_{1\le t,\hat{t}\le T} G_{t,\hat{t}} \left(\langle \mathbf{w}_t, D_{t,\hat{t}}^{+}\mathbf{w}_t\rangle + \langle \mathbf{w}_{\hat{t}}, D_{t,\hat{t}}^{+}\mathbf{w}_{\hat{t}}\rangle + \epsilon D_{t,\hat{t}}^{+}\right)$$
$$\text{s.t.} \quad \boldsymbol{\lambda}^{(t,\hat{t})} \in \mathcal{M}. \quad (5)$$

Algorithm 1 The KISAR algorithm

Input: $\{B_i, Y_i\}_{i=1}^{n}, G, d, \gamma$
Output: Predictors $\mathbf{W}$

1: Perform clustering, *e.g.*, k-means, on all the instances and obtain the mapping $\Phi(B)$ according to Eq.1.
2: Set $\boldsymbol{\lambda}_i^{(t,\hat{t})} = 1/d, \forall i = 1, \ldots, d, 1 \le t, \hat{t} \le T$.
3: Repeat until convergence:
   a) Fix $\{\boldsymbol{\lambda}^{(t,\hat{t})}\}_{t,\hat{t}=1}^{T}$, $\mathbf{W} \leftarrow$ solving Eq.6;
   b) Fix $\mathbf{W}$, $\{\boldsymbol{\lambda}^{(t,\hat{t})}\}_{t,\hat{t}=1}^{T} \leftarrow$ solving Eq.8.

Here, a small constant ϵ (*e.g.*, 10^{-3} in our experiments) is introduced to ensure the convergence (Argyriou *et al.* 2008).

Corollary 1 *Eq.5 is jointly-convex for both* $\mathbf{W}$ *and* $\{\boldsymbol{\lambda}^{(t,\hat{t})}\}_{t,\hat{t}=1}^{T}\}$.

Proof. The key is to show $\langle \mathbf{w}_t, D_{t,\hat{t}}^{+}\mathbf{w}_t\rangle$ is jointly-convex for $\mathbf{w}_t$ and $\boldsymbol{\lambda}^{(t,\hat{t})}$, whereas the proof can be found in (Boyd and Vandenberghe 2004). □

With Corollary 1 and according to (Bezdek and Hathaway 2003; Argyriou *et al.* 2008), alternating minimization algorithm can be applied to achieve a global solution of Eq.5. Specifically, when $\{\boldsymbol{\lambda}^{(t,\hat{t})}\}_{t,\hat{t}=1}^{T}$ is fixed, notice that all $\mathbf{w}_t$'s in Eq.5 are decoupled and thus $\mathbf{W}$ can be solved via multiple independent quadratic programming (QP) subproblems, *i.e.*,

$$\min_{\mathbf{w}_t} \gamma \sum_{i=1}^{n} \ell(y_{i,t}, \mathbf{w}_t'\phi(B_i)) + \langle \mathbf{w}_t, \sum_{1\le \hat{t}\le T} G_{t,\hat{t}} D_{t,\hat{t}}^{+}\mathbf{w}_t\rangle, \forall t. \quad (6)$$

Let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$ be the singular value decomposition (SVD) of $\sum_{1\le\hat{t}\le T} G_{t,\hat{t}} D_{t,\hat{t}}^{+}$. Denote $\bar{\mathbf{w}}_t = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{w}_t$ and $\bar{\phi}(B) = \mathbf{U}\boldsymbol{\Lambda}^{-1/2}\phi(B)$, and thus Eq.6 can be rewritten as:

$$\min_{\mathbf{w}_t} \quad \gamma \sum_{i=1}^{n} \ell(y_{i,t}, \bar{\mathbf{w}}_t'\bar{\phi}(B_i)) + \langle \bar{\mathbf{w}}_t, \bar{\mathbf{w}}_t\rangle, \forall t, \quad (7)$$

which is a standard SVM problem and thus Eq.7 can be addressed by state-of-the-art SVM solvers like LIBSVM (Hsieh *et al.* 2008) using SMO algorithm in an efficient manner. When $\mathbf{W}$ is fixed, according to (Argyriou *et al.* 2008), $\{\boldsymbol{\lambda}^{(t,\hat{t})}\}_{t,\hat{t}=1}^{T}$ can be solved via a closed-form solution, *i.e.*,

$$\boldsymbol{\lambda}^{(t,\hat{t})} = \left[\frac{\sqrt{w_{t,1}^2 + w_{\hat{t},1}^2 + \epsilon}}{Z}, \ldots, \frac{\sqrt{w_{t,d}^2 + w_{\hat{t},d}^2 + \epsilon}}{Z}\right], \quad (8)$$

where $Z = \|\mathbf{w}^{(t,\hat{t})}\|_{2,1} + d\epsilon$. The above procedures are executed iteratively until convergence. Algorithm 1 summarizes the KISAR algorithm.

Experiments

In this section, we first compare KISAR with state-of-the-art MIML algorithms on benchmark data sets. Then we evaluate KISAR on a real-world image annotation task. Finally, we study the pattern-label relations discovered by KISAR .

Four MIML algorithms are evaluated in the comparison: MIMLSVM (Zhou and Zhang 2006), MIMLBOOST (Zhou and Zhang 2006), MIMLKNN (Zhang 2010) and MIMLRBF (Zhang and Wang 2009). The codes of these algorithms are shared by their authors. In addition, to examine the label relations, KISAR is further compared with its two variants: KISARMINUS, a degenerated algorithm of KISAR which does not consider label relations (*i.e.*, the label relation matrix G is set to the identity matrix), and KISARALL, which considers all pairs of label relations (*i.e.*, G is set to the all-one matrix).

The compared MIML algorithms are set to the best parameters reported in the papers (Zhou and Zhang 2006; Zhang 2010; Zhang and Wang 2009). Specifically, for MIMLSVM, Gaussian kernel with width 0.2 is used and the number of clusters is set to 20% of the training bags; for MIMLKNN, the number of nearest neighbors and the number of citers are set to 10 and 20, respectively; for MIMLRBF, the fraction and scaling factors are set to 0.1 and 0.6, respectively. Due to the computational load of MIMLBOOST, the results of MIMLBOOST reported in (Zhou and Zhang 2006; Zhang and Wang 2009) are directly listed for comparison. As for KISAR and its two variants, the number of cluster d is fixed to 50% of the training bags and the parameter γ is selected from $\{10^{-2}, \ldots, 10^{3}\}$ by 10-fold cross validation on training sets.

The performances are evaluated with five popular multi-label measurements, *i.e.*, *hamming loss (h.l.)*, *one error (o.e.)*, *coverage (co.)*, *ranking loss (r.l.)* and *average precision (a.p.)*. Detailed descriptions about these measurements can be found in (Zhou *et al.* 2012; Schapire and Singer 2000).

Performance Comparison on MIML Benchmark

Text Categorization The MIML benchmark data set of text categorization is collected from the Reuters-21578 collection (Sebastiani 2002). In this data set, the seven most frequent categories are considered. There are $2,000$ documents in total, where around 15% documents are with multiple labels and the average number of labels per document is 1.15 ± 0.37. Each document is represented by a bag of instances by means of the sliding windows techniques (Andrews *et al.* 2003), where each instance corresponds to a text segment enclosed in a sliding window of size 50 (overlapped with 25 words). Each instance in the bags is represented as a 243-dimensional feature vector using the *Bag-of-Words* representation according to term frequency (Sebastiani 2002). The data set is publicly available.[2] For KISAR, without further domain knowledge, the entries of label relation matrix G is set to the concurrence of positive bags, *i.e.*,

$$G_{t,\hat{t}} = I(\sum\nolimits_{i=1}^{n} I(y_{i,t} = 1) I(y_{i,\hat{t}} = 1) > \theta).$$

Here, I is the identity function and θ is a threshold picked from $\{2^{-3}\theta_0, \ldots, 2^{3}\theta_0\}$ where θ_0 denotes the average concurrence between all label pairs, *i.e.*, $\theta_0 = \sum_{1 \le t, \hat{t} \le T} \sum_{i=1}^{n} I(y_{i,t} = 1) I(y_{i,\hat{t}} = 1) / T^2$. Five times ten-fold cross-validation (i.e., we repeat 10-fold cross validation for five times with different random data partitions) are conducted and the average performances are recorded.

Table 1: Performance comparison (mean±std.) on *text* data set. The best performance (paired t-tests at 95% significance level) and its comparable results are bolded. The ↓(↑) implies the smaller (larger), the better.

Algo	*Evaluation Metric*				
	h.l.$^{\downarrow}$	*o.e.*$^{\downarrow}$	*co.*$^{\downarrow}$	*r.l.*$^{\downarrow}$	*a.p.*$^{\uparrow}$
KISAR	**0.032 ±0.005**	**0.061 ±0.018**	**0.278 ±0.045**	**0.019 ±0.006**	**0.963 ±0.010**
KISAR MINUS	0.038 ±0.005	0.075 ±0.019	0.311 ±0.043	0.024 ±0.007	0.952 ±0.011
KISAR ALL	0.069 ±0.000	0.123 ±0.003	0.403 ±0.006	0.040 ±0.001	0.924 ±0.02
MIML SVM	0.044 ±0.006	0.105 ±0.024	0.373 ±0.054	0.034 ±0.008	0.934 ±0.014
MIML KNN	0.063 ±0.008	0.124 ±0.024	0.489 ±0.071	0.051 ±0.010	0.917 ±0.014
MIML RBF	0.061 ±0.008	0.123 ±0.024	0.418 ±0.071	0.042 ±0.010	0.924 ±0.017
MIML BOOST	0.053 ±0.009	0.107 ±0.022	0.417 ±0.047	0.039 ±0.007	0.930 ±0.012

Table 1 shows that KISAR obtains the best performance on all the measurements (the results of MIMLBOOST are from (Zhang and Wang 2009)). Paired t-tests at 95% significance level indicate that KISAR is significantly better than its two variants as well as MIMLSVM, MIMLKNN and MIMLRBF on all the five measurements.

Scene Classification The MIML benchmark data set of scene classification contains $2,000$ natural scene images with each image manually assigned by multiple labels from all possible class labels: *desert*, *mountains*, *sea*, *sunset* and *trees*. Over 22% of these images belong to more than one class and the average number of labels for each image is 1.24 ± 0.44. Each image is represented by a bag of nine 15-dimensional instances using the SBN image bag generator (Maron and Ratan 1998), where each instance corresponds to an image patch. The data set is publicly available.[3] Without further domain knowledge, the label relation matrix used in KISAR is set as the same as that in text categorization. Five times ten-fold cross-validation are conducted and average performances are recorded for comparison.

Table 2 shows that KISAR achieves the best performance on all measurements (the results of MIMLBOOST are from (Zhou and Zhang 2006)). Paired t-tests at 95% significance level indicate that KISAR is significantly better than its two variants as well as MIMLSVM, MIMLKNN and MIMLRBF on all the five measurements.

Performance Comparison on Image Annotation

In this experiments, a subset of the MSRA-MM database (Li *et al.* 2009a) of Microsoft Research Asia is used. All the la-

[2] http://lamda.nju.edu.cn/datacode/miml-text-data.htm

[3] http://lamda.nju.edu.cn/datacode/miml-image-data.htm

Table 2: Performance comparison (mean±std.) on *scene* data set. The best performance (paired t-tests at 95% significance level) and its comparable results are bolded. The ↓(↑) implies the smaller (larger), the better.

Algo	*Evaluation Metric*				
	h.l.$^{\downarrow}$	*o.e.*$^{\downarrow}$	*co.*$^{\downarrow}$	*r.l.*$^{\downarrow}$	*a.p.*$^{\uparrow}$
KISAR	**0.167 ±0.010**	**0.298 ±0.030**	**0.928 ±0.070**	**0.162 ±0.016**	**0.804 ±0.018**
KISAR MINUS	0.168 ±0.011	0.302 ±0.028	0.942 ±0.064	0.166 ±0.015	0.801 ±0.017
KISAR ALL	0.190 ±0.000	0.341 ±0.007	0.962 ±0.014	0.172 ±0.004	0.783 ±0.004
MIML SVM	0.184 ±0.014	0.338 ±0.036	1.039 ±0.075	0.190 ±0.017	0.776 ±0.020
MIML KNN	0.172 ±0.010	0.324 ±0.029	0.944 ±0.074	0.169 ±0.016	0.792 ±0.017
MIML RBF	0.169 ±0.011	0.310 ±0.031	0.950 ±0.069	0.169 ±0.017	0.797 ±0.018
MIML BOOST	0.189 ±0.007	0.335 ±0.021	0.947 ±0.056	0.172 ±0.011	0.785 ±0.012

bels are annotated by humans. The data set we used contains $1,605$ examples and thirty-eight class labels, *i.e.*, *'airplane'*, *'animal'*, *'baby'*, *'beach'*, *'bike'*, *'bird'*, *'boat'*, *'building'*, *'bus'*, *'candle'*, *'car'*, *'cat'*, *'cattle'*, *'cloud'*, *'desert'*, *'dog'*, *'dolphin'*, *'elephant'*, *'fire'*, *'fireworks'*, *'horse'*, *'ice'*, *'jungle'*, *'landscape'*, *'leaf'*, *'lightning'*, *'mountains'*, *'penguin'*, *'people'*, *'rock'*, *'sea'*, *'ship'*, *'sky'*, *'sun'*, *'swimming'*, *'water'*, *'waterfall'* and *'woman'*. Around 92% of these images are with more than one labels and there are at most eleven labels annotated to one example. The average number of labels for each image is 3.85 ± 1.75. Followed by (Wang *et al.* 2001), each image is represented as a bag of 6-dimensional instances based on image segmentation, where each instance corresponds to the cluster center of one segment and the number of segments is set to 16. Highly relevant labels, as shown in Table 3, are manually labeled by volunteers. For each trial, $1,400$ images are randomly selected for training and the remaining images are used for testing. Experiments are repeated for 10 times and the average performances are recorded.

Table 4 summarizes the results, where MIMLBOOST is not included since it does not return results within a reasonable time (24 hours in our experiments) in a singe trial. It can be seen that the performances of KISAR are quite good. Paired t-tests at 95% significance level show that KISAR is significantly better than KISARMINUS on all the measurements except that on *one error* and *hamming loss* there are no significant difference, whereas KISAR performs significantly better than all other compared methods on all the measurements.

Discovery of Pattern-Label Relation

Now we examine the pattern-label relations discovered by KISAR via intuitive illustrations (we are unaware of other effective ways for examining the pattern-label relations at present). Specifically, all the illustrated images are picked from the image annotation data. The key instances are identified as follows: For each label, the most important prototypes identified based on the rank of weights in linear predictors are recorded. Recall that each prototype corresponds to a cluster center, and thus, the instances within the clusters of the most important prototypes are realized as key instances triggering this label. For the sake of simplicity in displaying the relations, one key instance is picked for each label in each image.

Table 3: Related labels in image annotation.

airplane	*cloud*, *sky*
animal	*bird*, *cat*, *cattle*, *dog*, *dolphin*, *elephant*, *horse*, *penguin*
baby	*people*, *woman*
beach	*desert*, *sea*
boat	*ship*
building	*landscape*
bus	*car*
candle	*fire*, *sun*
cloud	*sky*
fire	*sun*
fireworks	*lightning*
jungle	*leaf*, *mountains*
leaf	*mountains*
people	*woman*
sea	*sun*, *swimming*, *water*, *waterfall*
swimming	*water*, *waterfall*
water	*waterfall*

Table 4: Performance comparison (mean±std.) on image annotation. The best performance (paired t-tests at 95% significance level) and its comparable results are bolded. The ↓(↑) implies the smaller (larger), the better.

Method	*Evaluation Metric*				
	h.l.$^{\downarrow}$	*o.e.*$^{\downarrow}$	*co.*$^{\downarrow}$	*r.l.*$^{\downarrow}$	*a.p.*$^{\uparrow}$
KISAR	**0.069 ±0.002**	**0.213 ±0.024**	**9.413 ±0.409**	**0.080 ±0.007**	**0.713 ±0.020**
KISAR MINUS	**0.069 ±0.003**	**0.212 ±0.024**	10.030 ±0.433	0.088 ±0.007	0.709 ±0.019
KISAR ALL	0.070 ±0.003	0.223 ±0.031	10.012 ±0.498	0.086 ±0.009	0.702 ±0.022
MIML SVM	0.077 ±0.002	0.263 ±0.021	14.647 ±0.657	0.145 ±0.014	0.634 ±0.027
MIML KNN	0.083 ±0.005	0.242 ±0.032	12.782 ±1.079	0.122 ±0.016	0.678 ±0.018
MIML RBF	0.073 ±0.003	0.242 ±0.037	12.264 ±0.784	0.110 ±0.017	0.686 ±0.021

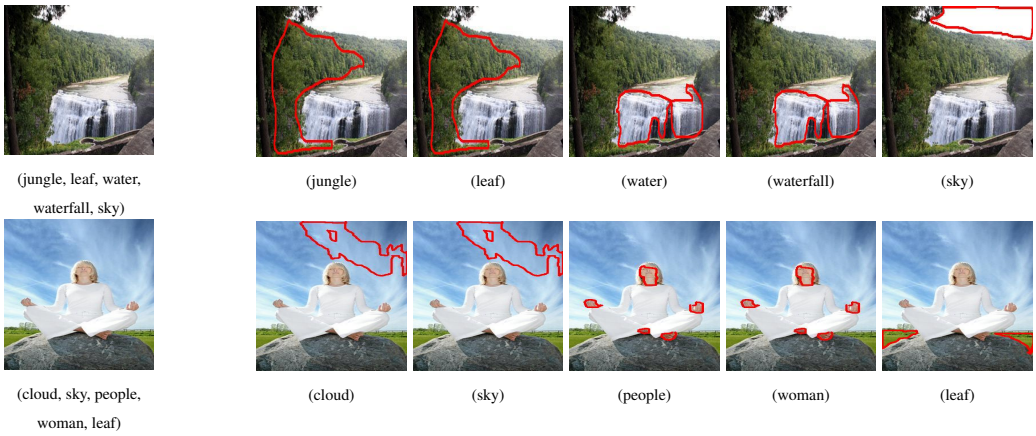

Figure 3: Pattern-label relations for images containing multiple pairs of highly relevant labels (above: {*'jungle', 'leaf'*} and {*'water','waterfall'*}, bottom: {*'cloud', 'sky'*} and {*'people','women'*}). The red contour highlights the identified key instances triggering the labels.

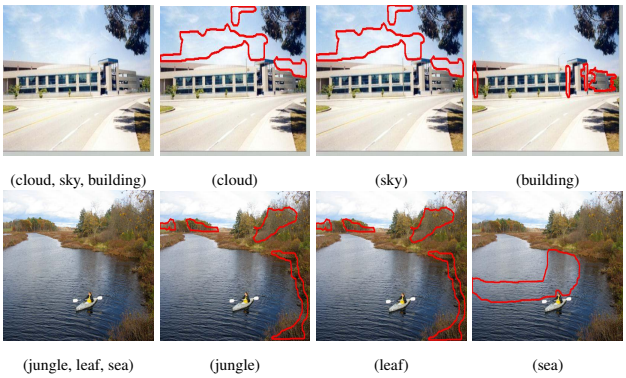

Figure 2: Pattern-label relations for images containing a single pair of highly relevant labels (above: {*'cloud', 'sky'*}, bottom: {*'jungle', 'leaf'*}). The red contour highlights the identified key instances triggering the labels.

We first study simpler cases where images contain a single pair of highly relevant labels. Figure 2 shows some original images and key instances. As can be seen, KISAR identifies reasonable patterns for labels. Specifically, for the first example, referring to Table 3, *sky* and *cloud* are relevant labels that are verified by their key instances; *sky* and *building* are not related, and this is also verified by their different key instances. Similar observations can be found for the second example. We further study more complicated cases where images contain multiple pairs of highly relevant labels simultaneously. Figure 3 shows some examples. As can be seen, KISAR is still capable of identifying reasonable patterns for the labels. These observations validate that MIML is capable of discovering relations between input patterns and output labels corresponding to different semantics.

Conclusion

In contrast to previous MIML studies that focused on improving generalization, in this paper, we propose the KISAR algorithm which is able to discover the relation between patterns in the input space and labels corresponding to different semantics in the output space. Although it has been pointed out before that the MIML framework offers the possibility of disclosing such relations (Zhou *et al.* 2012), to the best of our knowledge, none existing MIML algorithm was developed for this purpose. Our KISAR algorithm works based on the assumption that highly related labels generally share some common key instances. We get a convex formulation and provide an alternating optimization solution. Experimental results show that the predictive performances of KISAR are highly competitive than state-of-the-art MIML algorithms; more importantly, KISAR is capable of discovering some intuitively reasonable relations between input patterns and output labels.

There are many interesting further works. For example, our current proposal adopts a two-stage method where some useful information may be lost, whereas direct approaches like (Andrews *et al.* 2003; Li *et al.* 2009b) are worth trying in the future. Moreover, exclusive segmentations are employed in our work for input patterns, whereas overlapped segmentations with different scales of granularity might be more reasonable in many cases. This is an interesting issue to be studied in the future.

References

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press, Cambridge, MA, 2003.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

J.C. Bezdek and R.J. Hathaway. Convergence of alternating optimization. *Neural, Parallel & Scientific Computations*, 11(4):351–368, 2003.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Y. Chen and J. Z. Wang. Image categorization by learning

and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

D. L. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006.

D.M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

C. J. Hsieh, K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th International Conference on Machine Learning*, pages 408–415, Helsinki, Finland, 2008.

S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th International Conference on Machine Learning*, pages 457–464, Montreal, Canada, 2009.

R. Jin, S. Wang, and Z.-H. Zhou. Learning a distance metric from multi-instance multi-label data. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 896–902, Miami, FL, 2009.

H. Li, M. Wang, and X.S. Hua. MSRA-MM 2.0: A large-scale web multimedia dataset. In *Proceedings of the International Conference on Data Mining Workshops*, pages 164–169, 2009.

Y.-F. Li, J. T. Kwok, I. W. Tsang, and Z.-H. Zhou. A convex method for locating regions of interest with multi-instance learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 15–30, Bled, Slovenia, 2009.

A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. In *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, pages 8–14, 1998.

J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2, 1-norm minimization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 339–348, Montreal, Canada, 2009.

J. Luo and F. Orabona. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems 23*, pages 1504–1512. MIT Press, Cambridge, MA, 2010.

O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems 10*, pages 570–576. MIT Press, Cambridge, MA, 1998.

O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of 15th International Conference on Machine Learning*, pages 341–349, Madison, WI, 1998.

Y. Nesterov and I. U. E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2004.

N. Nguyen. A new SVM approach to multi-instance multi-label learning. In *Proceeding of the 10th IEEE International Conference on Data Mining*, pages 384–392, Sydney, Australia, 2010.

M. Omologo, P. Svaizer, and M. Matassoni. Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, 25(1-3):75–95, 1998.

R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168, 2000.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

E. M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the 12th Text Retrieval Conference*, pages 54–68, 2003.

J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Learning*, 23(9):947–963, 2001.

S.-H. Yang, H. Zha, and B.-G. Hu. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Advances in Neural Information Processing Systems 22*, pages 2143–2150. MIT Press, Cambridge, MA, 2009.

Z.J. Zha, X.S. Hua, T. Mei, J. Wang, G.J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AL, 2008.

M.-L. Zhang and Z.-J. Wang. MIMLRBF: RBF neural networks for multi-instance multi-label learning. *Neurocomputing*, 72(16-18):3951–3956, 2009.

M.-L. Zhang and Z.-H. Zhou. M3MIML: A maximum margin method for multi-instance multi-label learning. In *Proceeding of the 9th IEEE International Conference on Data Mining*, pages 688–697, Miami, FL, 2009.

M.-L. Zhang. A k-nearest neighbor based multi-instance multi-label learning algorithm. In *Proceeding of the 22nd International Conference on Tools with Artificial Intelligence*, pages 207–212, Arras, France, 2010.

Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *Advances in Neural Information Processing Systems 19*, pages 1609–1616. MIT Press, Cambridge, MA, 2006.

Z.-H. Zhou and M.-L. Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems*, 11(2):155–170, 2007.

Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.