Multi-Instance Learning with Discriminative Bag Mapping

Jia Wu, Shirui Pan, Xingquan Zhu, *Senior Member, IEEE*, Chengqi Zhang, *Senior Member, IEEE*, and Xindong Wu, *Fellow, IEEE*

Abstract—Multi-instance learning (MIL) is a useful tool for tackling labeling ambiguity in learning because it allows a bag of instances to share one label. Bag mapping transforms a bag into a single instance in a new space via instance selection and has drawn significant attention recently. To date, most existing work is based on the original space, using all instances inside each bag for bag mapping, and the selected instances are not directly tied to an MIL objective. As a result, it is difficult to guarantee the distinguishing capacity of the selected instances in the new bag mapping space. In this paper, we propose a discriminative mapping approach for multi-instance learning (MILDM) that aims to identify the best instances to directly distinguish bags in the new mapping space. Accordingly, each instance bag can be mapped using the selected instances to a new feature space, and hence any generic learning algorithm, such as an instance-based learning algorithm, can be used to derive learning models for multi-instance classification. Experiments and comparisons on eight different types of real-world learning tasks (including 14 data sets) demonstrate that MILDM outperforms the state-of-the-art bag mapping multi-instance learning approaches. Results also confirm that MILDM achieves balanced performance between runtime efficiency and classification effectiveness.

Index Terms—Multi-instance learning, instance selection, bag mapping, classification

1 Introduction

In generic supervised learning, each training sample is an instance associated with a class label (e.g., positive or negative), as shown in Fig. 1. By contrast, in multi-instance learning (MIL), each training object is a bag that contains a number of instances. A label is assigned to the bag, but not to the individual instances, under the constraint that all the instances in a negative bag are negative and at least one instance in a positive bag should be positive. This is known as the MIL assumption and is illustrated in Fig. 2.

Multi-instance learning was initially investigated by Dietterich et al. [1] to capture the unstable characteristics and complex behaviors that occur during drug activity prediction. Because molecular activity can vary significantly or show different behaviors in response to changing environments, the feature values of a specific molecule can change when observed in different experiments. An efficient way to accommodate such changing behaviors is to represent the molecule as a bag of instances with each instance representing one observed behavior of the molecule. If a molecule shows positive/interested behavior in a particular experiment, the bag is labeled as positive but, if no positive behavior is demonstrated in any experiment, the bag is labeled as negative. MIL's performance in this unique setting suggested its effectiveness for accommodating labeling ambiguity in other real-world applications. For example, in content-based image classification, each region of the image can be regarded as an instance. An image can, therefore, be represented as a bag containing a number of instances. If one region of the image contains an object of interest (e.g., an animal), the image/bag is labeled as positive [2], and MIL can be used to identify bags containing objects of interest [3], [4]. Such an approach has also been used for text categorization [5], sign language recognition [6], saliency detection [7], graph mining [8], [9], [10], web recommendation [11], and so on.

Existing MIL solutions [12], [13], [14], [15] can be roughly divided into two categories: (a) updating a generic learning algorithm to tackle label ambiguity problems, or (b) developing a learning paradigm specifically for multiple instance learning. However, the performance of the above methods deteriorates when there are a large number of instances in a bag [3]. Using content-based image classification, again, as an example, the total number of instances in a bag could be extremely large if the image contains many regions. However, in reality, different regions/instances in a bag may make different contributions to image classification and the more informative the instances, the more information can be provided to learning tasks. In this scenario, selecting the most informative instances in each bag becomes a challenging problem for MIL [16].

One common approach is to convert multi-instance learning (bag learning) into a more traditional form of supervised learning (single-instance learning). For example, one might propagate the bag label to the instances inside the bag so that a propositional classifier can be learned for bag classification [17], [18]. However, transmitting the label of a positive bag

- *J. Wu is with the Department of Computing, Faculty of Science and Engineering, Macquarie University, Sydney, NSW 2109, Australia. E-mail: jia.wu@mq.edu.au.*
- *S. Pan and C. Zhang are with Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia. E-mail: {shirui.pan, chengqi.zhang}@uts.edu.au.*
- *X. Zhu is with the Department of Computer and Electrical Engineering & Computer Science, Florida Atlantic University, Boca Raton, FL 33431. E-mail: xzhu3@fau.edu.*
- *X. Wu is with the School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, LA 70504. E-mail: xwu@louisiana.edu.*

Manuscript received 6 Apr. 2016; revised 10 July 2017; accepted 11 July 2017. Date of publication 1 Jan. 2018; date of current version 27 Apr. 2018. (Corresponding author: Chengqi Zhang.) Recommended for acceptance by Y. Chang. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2017.2788430

1041-4347 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

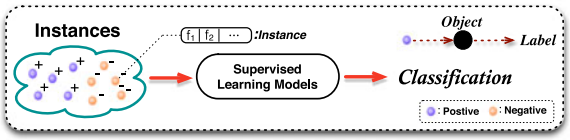


Fig. 1. Traditional supervised learning: the labels (i.e., + or −) are available for each instance (i.e., ball). The *object* for classification is an instance.

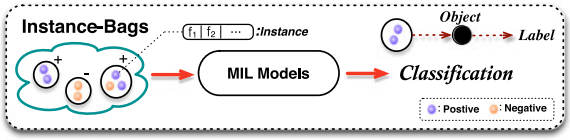


Fig. 2. Multi-instance learning: each bag (i.e., circle) consists of several instances (i.e., balls) and labels (i.e., + or −) are only available for bags. The *object* for classification is an instance bag.

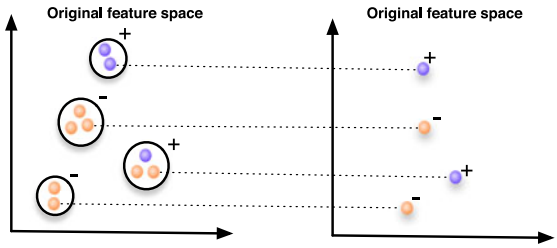


Fig. 3. Non-bag mapping approaches: Each bag is represented by an instance or uses the mean value of all instances inside the bag in the original feature space.

would assign all the negative instances inside with incorrect class labels. Alternatively, one instance could be used to represent each bag based on its statistic properties, known as bag representation. In [19] three different types of summarization approaches to bag representation were proposed—arithmetic mean, geometric mean, and minimax. Wu et al. [20] proposed a different method based on the distribution of the negative bags and, in doing so, transformed all bags into a set of instances in the same feature space, as shown in Fig. 3. Although this type of single-instance representation algorithm works reasonably well, it does discard most of the instance information in each bag.

Bag mapping with instance selection is another proposed approach, which represents each bag in a new feature space, as shown in Fig. 4. Chen et al. [3] proposed an embedding instance selection method that maps each bag into a new feature space created from a hidden instance set, i.e., an intermediate instance pool (IIP) constructed from a training bag of instances. Following this IIP-based bag mapping strategy, Fu et al. [16] proposed another bag mapping method that selects a subset of instances for bag-level feature computation using the distribution of negative instances. The main difference between these two methods is the construction of the IIP. The former approach chooses all the training instances as the IIP, while the latter approach selects a subset of the instance that is most likely to be positive from each positive bag according to the distribution of the negative instances. Either way, a good instance selection method may lead to better performance. According to the above observation, Hong et al. [2] proposed selecting all the instances from positive bags and the clustered instances in negative bags as the IIP. The bag mapping methods that rely on instance selection are able to prune the instance space; however, it may be difficult to distinguish between the instances in the new bag mapping space. Therefore,

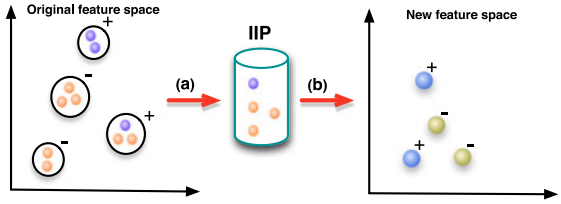


Fig. 4. Traditional bag mapping based on instance selection: Each bag is represented by an instance in the new feature space via an intermediate instance pool (IIP).

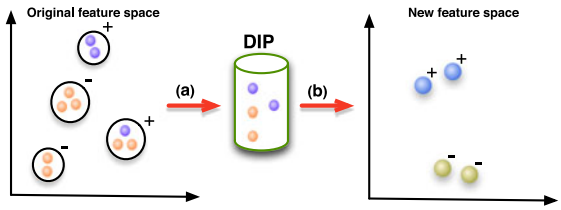


Fig. 5. Discriminative bag mapping: The informative instances are selected as a discriminative instance pool to ensure that the bags in the new mapping space can be easily distinguished.

designing efficient selection and instance pruning techniques for discriminative bag mapping is important.

In this paper, we propose a direct discriminative mapping approach for multi-instance learning (MILDM) that aims to identify the instances that will make the bags maximally distinguishable in the new mapping space, as shown in Fig. 5. Experiments and comparisons on eight different types of real-world learning tasks (including drug activity prediction, content-based image classification, train bound challenge, mutagenicity prediction, scientific publication retrieval, online product evaluation, newsgroup categorization, and web index recommendation) confirm the effectiveness of the proposed design. The contributions of this paper are threefold.

- An instance evaluation criterion based on a given bag is proposed as the instance pruning criterion.
- A discriminative bag mapping framework is proposed for multi-instance learning.
- Eight various learning tasks (including 14 data sets) are used to validate the generality of the MILDM.

The rest of the paper is organized as follows: in Section 2, we review related work on instance selection-based MIL. Section 3 outlines the proposed MILDM framework, followed by experiments in Section 4. Section 5 discusses the properties of the proposed MILDM, and we conclude the paper in Section 6.

2 Related Work

Multiple-instance learning is a variation of supervised learning [1]. Many real-world applications can be considered as MIL problems, and a variety of MIL approaches [12], [13], [14], [15], [21] exist to solve different MIL applications. Such algorithms can be categorized into two major groups: *upgraded single-instance learners and specifically designed MIL algorithms*. The former learners are an adaption of existing single-instance learning algorithms to support multi-instance learning. Lazy learning citation-KNN and Bayesian-KNN [22] extend the k-nearest neighbor algorithm (KNN) [23] to multi-instance settings. Other approaches include: tree-based multi-instance learning [24], multi-instance rule-based learning mi-DS algorithm [25], multi-instance kernel machine MISMO [26], multi-instance logistic learning MILR [27],

multi-instance ensemble learning MIBoost [28], and multi-instance bag dissimilarity-based learning [29].

Specifically designed MIL algorithms use bag constraints to reorganize the instances inside each bag into specific formats for learning. Axis parallel hyper-rectangles, proposed by Dietterich et al. [1], is an early approach of this type. Diverse density (DD) [30] searches for a point in the feature space by maximizing the diverse density function that measures the co-occurrence of similar instances in different positive bags. MIEMDD [31] is an improved multi-instance DD approach that can convert a multi-instance framework into a traditional single instance problem by employing the EM strategy. MIOptimalBall [32] builds an optimal ball to ensure that at least one instance in the positive bags is inside the ball and all the negative instances are outside the ball. Zhang et al. [33] proposed a novel multi-instance learning framework from multiple information sources. Xiao et al. [34] proposed a similarity-based classification framework for multiple instance learning.

Amores [14] discussed vocabulary-based methods, where a list of vocabulary stores information about all the classes of instances in the training set, and this information is used to first classify the instances in a new bag, then perform the embedding of this bag. Vocabulary-based approaches comprise four major groups: 1) *histogram-based methods* use a function that maps each bag into a histogram where each bin counts how many instances fall into a specific class of the vocabulary [13], [35], [36], [37]; 2) *distance-based methods* map each bag into a vector space by providing the lowest distance from a special class to any instance in the bag [38], [39], [40], [41]. MILES [3] and MILIS [16] both belong to this category; 3) *attribute-based methods* include a mapping function that returns a vector which is a concatenation of the sub-vectors that summarize the attributes of the instances that match a special class [42]; and 4) *vocabularies of bags-based methods* form a vocabulary from the classes of bags not the instances [43].

In reality, one potential problem that reduces the performance of the above approaches is that learning usually has to contend with a large number of instances for even moderate-sized data sets [3]. In these cases, selecting the most informative instances to represent each bag becomes a challenging problem [16]. A novel approach solving this issue is to use the selected instances from the bags to convert the MIL problem into a standard single-instance learning task [3], [16], [19], [20]. We call these methods "instance selection-based MIL", and they can be divided into two categories: non-bag mapping approaches and bag mapping approaches.

2.1 Instance Selection-Based Non-Bag Mapping

The basic idea of non-bag mapping methods is to choose one or multiple instances from each bag to represent the whole bag. An intuitive method is to directly propagate the bag label to all the instances inside the bag (i.e., using all instances to represent the bag) [18]. Three other commonly used non-bag mapping models include arithmetic mean, geometric mean and max-min [19]. The first two models are based on the assumption that each individual instance within a bag contributes independently and equally to the bag label. The arithmetic mean model simply calculates the arithmetic mean of the instances for each bag, while the geometric mean model calculates the geometric mean of the instances. This type of simple non-bag mapping strategy was used in [44]. The max-min model records both the minimum and maximum values of each dimension for every bag [45]. A further method, based on the distribution of the negative bags, has also been proposed as a non-bag mapping MIL method [20]. After choosing the representative instances from each bag, the MIL problem is converted into a generic supervised learning problem so that a conventional classifier can be applied to the new instances for learning. In reality, because most of the information about the instances in a bag are discarded, non-bag mapping methods tend to face the challenge of information loss and a deterioration in classification performance.

2.2 Instance Selection-Based Bag Mapping

The fundamental idea of bag mapping approaches is to choose a set of instance prototypes, i.e., an IIP, to map each bag into a new feature space. Two representative methods are MILES [3] and MILIS [16]. MILES does not define an explicit mechanism for instance prototype selection because the IIP is composed of all the instances in the training bags. Once the IIP is formed, MILES maps each bag into a feature space defined by the IIP using a bag-instance similarity measure. However, MILES might potentially map multi-instance learning into a high-dimensionality problem, because the dimensions of the mapping feature space depend on the size of IIP, i.e., the number of instances in training bags. To address this issue, MILIS selects only one instance from each positive bag to prune the instance space. The instance that is most likely to be positive in each positive bag is selected for the IIP, and this is determined by the likelihood of whether an instance is positive using the distributions of all the instances in negative bags. Once constructed, the IIP is used to map each bag into a new bag-level feature space so that a traditional classifier can be directly employed for further learning. A further type of IIP construction that consists of all the instances within all positive bags along with the clustering centers of instances in negative bags was proposed by Hong et al. [2]. However, because IIP instance selection is not directly tied to the underlying MIL learning problem, it is difficult to guarantee that the selected instances will be distinguishable from each other in the new bag mapping space.

3 MIL with Discriminative Bag Mapping

3.1 Preliminaries and Overall Framework

A bag B_i contains a number of instances, in which $x_{i,j}$ denotes the jth instance in the ith bag. The class label of the bag B_i is denoted by $y_i = \mathcal{Y}$, with $\mathcal{Y} = \{-1, +1\}$. The collection of all bag sets is denoted by $\mathcal{B}$.

In the training procedure, all bags B_i are transformed into B_i^{ϕ}, a single instance in a new feature space, using the discriminative instance pool (DIP), denoted as $\mathcal{P}$. $B_i^{\phi} = [s(B_i, x_1^{\phi}), \ldots, s(B_i, x_m^{\phi})]$, where $s(B_i, x_k^{\phi})$ denotes the similarity between bag B_i and the kth instance x_k^{ϕ} as determined by the instance candidate $x_k^{\phi} \in \mathcal{P}$. A transitional supervised learning classifier, i.e., instance-based learning algorithm IB1, is then trained on the instances in the new feature space. In the testing phase, we first map each test bag into a new instance based on the DIP obtained in the training process. Then, the trained learning classifier is used to predict the final class label. The most important part of this process is finding the optimal DIP for bag mapping.

3.2 DIP Optimization

Given $\mathcal{B}$ with n bags, and an instance set $\mathcal{X}$ of size p collected from all bags in $\mathcal{B}$, our objective is to find a subset

$\mathcal{P} \subseteq \mathcal{X}$ using an instance selection matrix $\mathcal{I}_\mathcal{P}$ (a diagonal matrix, $diag(\mathcal{I}_\mathcal{P}) = \mathbf{d}(\mathcal{P})$), where $\mathbf{d}(\mathcal{P})$ is an indicator vector, if $x_i \in \mathcal{P}$, $\mathbf{d}(\mathcal{P})_i = 1$, or otherwise 0.

Accordingly, we define $\mathcal{J}(\mathcal{P})$ as an instance evaluation function to measure $\mathcal{P}$ as follows:

$$\mathcal{P}_* = \arg\max_{\mathcal{P}\subseteq\mathcal{X}} \mathcal{J}(\mathcal{P}) \quad s.t.\ |\mathcal{P}| = m, \tag{1}$$

where $|\cdot|$ denotes the cardinality of the instance set, and m is the number of instances to be selected from $\mathcal{X}$ (i.e., the size of the DIP). The objective function in Eq. (1) states that the instances selected for MIL $\mathcal{P}_*$ should be maximally discriminative in the new mapping space.

3.3 Discriminative Instance Pool Evaluation Criteria

To obtain the DIP with the discriminative power, we impose a rule that the optimal DIP should have the following properties: (a) *bag mapping must-link*. Because each bag B_i is associated with a class label (positive or negative), the selected DIP should ensure that the bags B_i^ϕ with the same label are similar to each other in the mapping space; and (b) *bag mapping cannot-link*. Bags with different class labels in the mapping space should represent the disparity between them. Accordingly, the DIP evaluation criteria can be measured as

$$\mathcal{J}(\mathcal{P}) = \frac{1}{2}\sum_{i,j} K_\mathcal{P}\Big(B_i, B_j\Big) Q_{i,j}, \tag{2}$$

where $K_\mathcal{P}(B_i, B_j)$ denotes the distance between two bags B_i and B_j in the new mapping space ϕ based on the DIP $\mathcal{P}$. Along with matrix Q embedding the class label information, $\mathcal{J}(\mathcal{P})$ can represent level of discriminativeness in the mapping space. More specifically, $K_\mathcal{P}(B_i, B_j)$ can be formulated as

$$K_\mathcal{P}(B_i, B_j) = \left\|\mathcal{I}_\mathcal{P} B_i^{\phi_x} - \mathcal{I}_\mathcal{P} B_j^{\phi_x}\right\|^2, \tag{3}$$

where $B_i^{\phi_x}$, denoted in a similar way to B_i^ϕ, uses all the instances as the mapping instance pool. By defining the label embedding matrix Q as

$$Q_{i,j} = \begin{cases} -1/|A|, & y_i y_j = 1 \\ 1/|B|, & y_i y_j = -1, \end{cases} \tag{4}$$

where $A = \{(i,j)|y_i y_j = 1\}$ denotes the *bag mapping must-link* pairwise bag constraint sets and $B = \{(i,j)|y_i y_j = -1\}$ denotes the *bag mapping cannot-link* pairwise sets. We can rewrite $\mathcal{J}(\mathcal{P})$ in Eq. (1) as follows:

$$\begin{aligned}
\mathcal{J}(\mathcal{P}) &= \frac{1}{2}\sum_{i,j} \left\|\mathcal{I}_\mathcal{P} B_i^{\phi_x} - \mathcal{I}_\mathcal{P} B_j^{\phi_x}\right\|^2 Q_{i,j} \\
&= \sum_{i,j} (B_i^{\phi_x})^\top \mathcal{I}_\mathcal{P}^\top \mathcal{I}_\mathcal{P} B_i^{\phi_x} Q_{i,j} - \sum_{i,j} (B_i^{\phi_x})^\top \mathcal{I}_\mathcal{P}^\top \mathcal{I}_\mathcal{P} B_j^{\phi_x} Q_{i,j} \\
&= \sum_{i} (B_i^{\phi_x})^\top \mathcal{I}_\mathcal{P}^\top \mathcal{I}_\mathcal{P} B_i^{\phi_x} \sum_{j} Q_{i,j} - \sum_{i,j} (B_i^{\phi_x})^\top \mathcal{I}_\mathcal{P}^\top \mathcal{I}_\mathcal{P} B_j^{\phi_x} Q_{i,j} \\
&= \sum_{i} (B_i^{\phi_x})^\top \mathcal{I}_\mathcal{P}^\top \mathcal{I}_\mathcal{P} B_i^{\phi_x} D_{i,i} - \sum_{i,j} (B_i^{\phi_x})^\top \mathcal{I}_\mathcal{P}^\top \mathcal{I}_\mathcal{P} B_j^{\phi_x} Q_{i,j} \\
&= tr(\mathcal{I}_\mathcal{P}^\top \mathcal{X}_\phi D \mathcal{X}_\phi^\top \mathcal{I}_\mathcal{P}) - tr(\mathcal{I}_\mathcal{P}^\top \mathcal{X}_\phi Q \mathcal{X}_\phi^\top \mathcal{I}_\mathcal{P}) \\
&= tr(\mathcal{I}_\mathcal{P}^\top \mathcal{X}_\phi (D - Q) \mathcal{X}_\phi^\top \mathcal{I}_\mathcal{P}) \\
&= tr(\mathcal{I}_\mathcal{P}^\top \mathcal{X}_\phi L \mathcal{X}_\phi^\top \mathcal{I}_\mathcal{P}) \\
&= \sum_{x_k^\phi \in \mathcal{P}} \boldsymbol{\phi}_k^\top L \boldsymbol{\phi}_k,
\end{aligned} \tag{5}$$

where $tr(\cdot)$ denotes the matrix trace operator, $\mathcal{X}_\phi = [B_1^{\phi_x}, \ldots, B_n^{\phi_x}] = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p]^\top \in \{\mathbb{R}\}^{p\times n}$, with n denoting the size of bag. D, as a diagonal matrix, is generated from Q, where $D_{i,i} = \sum_j Q_{ij}$. L is a Laplacian matrix generalized from Q, denoted as $L = [L_{i,j}]^{n\times n} = D - Q$. By using function $f(x_k^\phi, L)$ to denote $\boldsymbol{\phi}_k^\top L \boldsymbol{\phi}_k$, the original optimization problem in Eq. (1) can be translated to maximize the sum of $f(x_k^\phi, L)$ with respect to optimal instance mapping set $\mathcal{P}$ as

$$\max_{\mathcal{P}\subseteq\mathcal{X}} \sum_{x_k^\phi \in \mathcal{P}} f(x_k^\phi, L) \quad s.t.\ |\mathcal{P}| = m. \tag{6}$$

To find the optimal instance set $\mathcal{P}$ that maximizes the criterion defined in Eq. (1), we can calculate the score of each instance (i.e., $\boldsymbol{\phi}_k^\top L \boldsymbol{\phi}_k$ in $\mathcal{X}$, then collect the top-m instances as the final DIP.

Algorithm 1. DIP: Discriminative Instance Pool

Input:
Training bag data set $\mathcal{B}$;
The instance set $\mathcal{X}$ collected from $\mathcal{B}$;
The number of selected mapping instance m;
Output:
$\mathcal{P} = \{p_1, \ldots, p_m\}$: A set of mapping instances;
1: $\mathcal{P} = \emptyset, \tau = 0$;
2: $Q \leftarrow$ Apply all bag labels in $\mathcal{B}$ to obtain the label embedding matrix via Eq. (4).
3: $L \leftarrow$ Apply Q to obtain the corresponding Laplacian matrix.
4: **for** each instance x_k in the $\mathcal{X}$ **do**
5: $f(x_k, L) \leftarrow$ Apply Eq. (6) to measure the score.
6: **if** $|\mathcal{P}| \leq m$ or $f(x_k, L) > \tau$, **then**
7: $\mathcal{P} \leftarrow \mathcal{P} \bigcup x_k$;
8: **if** $|\mathcal{P}| \geq m$, **then**
9: $\mathcal{P} \leftarrow \mathcal{P}/\arg\min_{x_k\in\mathcal{P}} f(x_k, L)$;
10: $\tau = \min_{x_k\in\mathcal{P}} f(x_k, L)$;
11: **end for**
12: **return** $\mathcal{P}$;

Algorithm 1 sets out the proposed DIP exploration approach. It starts with an empty instance set $\mathcal{P} = \emptyset$ and a minimum score $\tau = 0$ (line 1). The label embedding matrix Q is calculated first, along with its corresponding Laplacian matrix L (lines 2-3). Then, each instance $x_k \in \mathcal{X}$ is enumerated, and its discriminative score $f(x_k, L)$ is calculated based on matrix L, which embeds the label information. If $f(x_k, L)$ is greater than the minimum discriminative score in $\mathcal{P}$ as τ, or the size of $\mathcal{P}$ is less than m (i.e., $\mathcal{P}$ is not full), x_k is selected as one of the items for $\mathcal{P}$ (lines 6-7). Otherwise, if $\mathcal{P}$ overflows, the instance $\arg\min_{x_k\in\mathcal{P}} f(x_k, L)$ with the smallest discriminative score is removed from $\mathcal{P}$ to maintain its size (lines 8-9). Subsequently, the minimum discriminative score τ in $\mathcal{P}$ is updated (line 10). The loop continues until the final optimal discriminative instance pool $\mathcal{P}$ is derived.

3.4 Bag Mapping via Discriminative Instance Pool

Once the DIP has been constructed using selected instances, each bag needs to be mapped to a single instance in the new space. Given a DIP $\mathcal{P}$ with m instances, B_i can be mapped to a single instance $B_i^\phi = [s(B_i, x_1^\phi), \ldots, s(B_i, x_m^\phi)]$, with $s(B_i, x_k^\phi)$ denoting the similarity between the bag B_i and the kth instance x_k^ϕ as

$$s(B_i, x_k^\phi) = \max_{x_{i,j}\in B_i} exp(-\|x_{i,j} - x_k^\phi\|^2/\sigma^2), \tag{7}$$

where $x_{i,j}$ is the jth instance in the ith bag B_i, and σ is a predefined scaling factor. After each $B_i \in \mathcal{B}$ is mapped to B_i^{ϕ} based on the optimal DIP, any kind of single-instance learner (e.g., KNN) can be applied without constraint.

We design two types of discriminative bag mapping approaches.

3.4.1 Global Discriminative Bag Mapping

This type of bag mapping methods calculate the score of each instance in all bags and select the top-m instances as the DIP.

- *aMILGDM* uses all the training bags to generate the global DIP.
- *pMILGDM* only uses the positive bags.

3.4.2 Local Discriminative Bag Mapping

Local discriminative bag mapping approaches evaluate every instance inside the bag and select one instance with the highest discriminative score.

- *aMILLDM* selects the most discriminative instance from each bag to obtain the local DIP.
- *pMILLDM* chooses the instance with the highest discriminative score from each positive bag.

In short, the global MILDM, aMILGDM or pMILGDM, measures the discriminative power of the instances across the bags, while the local MILDM, aMILLDM or pMILLDM, compares the discriminative scores inside each individual bag. For global MILDM, the instances in the DIP may come from the same bag, so only a few bags might contribute to the learning procedure. By contrast, when using the local MILDM, the instances in the DIP are sourced from different bags. For simplicity, we also use pMILDM to denote the bag mapping approach that only evaluates positive bags (pMILLDM or pMILGDM), and aMILDM to denote the approach that evaluates all bags (aMILLDM or aMILGDM).

4 Experiments

4.1 Experimental Settings

The DIP, once constructed, can be used to map a bag to an instance by propagating the bag's label to the newly mapped instance. In this way, any supervised learning approach could be applied to support MIL classification. In our experiments, we use the instance-based learning algorithm IB1. In keeping with [7], [29], [46], [47], we have used the F-measure and area under ROC curve (AUC) as the evaluation metrics to validate the effectiveness of the proposed MILDM. The F-measure $= 2 \times P \times R/(P + R)$ combines recall R and precision P. AUC performance is calculated as $E = [P_0 - t_0(t_0 + 1)]/t_0t_1$, where t_0 and t_1 are the number of negative and positive instances, respectively. $P_0 = \sum r_i$, with r_i denoting the rank of the ith negative instance in the ranked list. All reported results are obtained through ten times 10-fold cross validation. The scaling parameter σ^2 is set to $8 * 10^5$, in keeping with previous works [3], [18]. The m values for the instance selection-based bag mapping approaches are derived as: the number of positive bags for pMILGDM, pMILLDM, and pMILIS; the number of all bags for aMILGDM, aMILLDM, and aMILIS; the number of all instances for MILES; and the number of all instances in positive bags plus the number of negative clusters for MILFM (in [2], this cluster number is set as the number of positive bags). All experiments are carried out on a Linux cluster node with an Intel(R) Xeon(R) @3.33 GHZ CPU and 3 GB fixed memory.

4.2 Baseline Methods

We use the following instance selection-based MIL baseline approaches from both non-bag mapping and bag mapping perspectives for comparison.

4.2.1 Non-Bag Mapping Instance Selection Approaches

In these approaches, a bag is directly represented by an instance or multiple instances inside the bag in the original feature space.

1. *MILMR* uses the mean of all instances inside each bag as the bag representation [19], [48].
2. *MILWA* propagates the bag label to all the instances inside the bag as the bag representation [17], [18].
3. *MILIR* uses the distribution of the negative bags to select one instance to represent the bag [20].

4.2.2 Bag Mapping Instance Selection Approaches

In bag mapping approaches, each bag is mapped into a single instance in the new feature space using an IIP constructed from the training bags.

4. *MILES* maps each bag into a feature space using all the training instances for the IIP via a bag-instance similarity measure [3].
5. *pMILIS* applies kernel density estimation (KDE) to select one instance from each positive training bag for the IIP, which is then used for further bag mapping [16].
6. *aMILIS* selects one instance from all the training bags for the IIP (i.e., the most positive instance or the least negative instance is selected from each positive bag and negative bag, respectively), and the instance selection strategy is the same as MILIS [16].
7. *MILFM* uses the all the instances in the positive bags and the clustered instances in the negative bags for the IIP [2].

4.3 Experimental Data Sets

Eight types of learning tasks across 14 data sets are used to validate MILDM. Table 1 shows the statistics for each data set. The original data sets for drug activity perdition, content-based image classification, newsgroup categorization, and web index recommendation tasks can be found at http://www.miproblems.org. The data sets for the train bound challenge, and the mutagenicity prediction task are available online at http://www.cs.waikato.ac.nz/~eibe/multi_instance/. The scientific publication retrieval and online product evaluation MIL data are available at http://web.science.mq.edu.au/~jiawu/data/MIL/TKDE18.DATA.zip. In the following, we briefly explain the domain knowledge of each data set.

4.3.1 Drug Activity Prediction Data

The objective of drug activity prediction is to predict the potency of the drug molecules on certain disease states. The

TABLE 1
Details of the Benchmark Data Sets

Data set	Pos.bags	Neg.bags	Atts	Insts	Avg/bag	Min/bag	Max/bag
Musk1	47	45	166	476	5	2	40
Musk2	39	63	166	6,598	64	1	1,044
Elephant	100	100	230	1,391	6	2	13
Tiger	100	100	230	1,220	6	1	13
EastWest	10	10	24	213	10	4	16
WestEast	10	10	24	213	10	4	16
Atom	125	63	10	1,618	8	5	15
Bond	125	63	16	3,995	21	8	40
AICV	100	100	4,497	1,151	5	1	10
Food	200	200	1,517	2,097	5	1	10
News.rm	50	50	200	4,730	47	22	73
News.tpm	50	50	200	3,376	33	15	55
Web7	54	59	6,450	3,423	30	4	200
Web8	55	58	5,999	3,423	30	4	200

data sets consist of descriptions of molecules. Each molecule is represented as a bag. Low-energy shapes, or conformations of the molecule, are the instances in the bag, as shown in Fig. 6. Because the bonds in molecules can rotate, each molecule can exhibit many different shapes. The Musk data sets (Musk1 and Musk2) [1] are the benchmark drug activity prediction data used for MIL. In Musk2, all the low-energy conformations of the molecules are used to generate the conformations. By contrast, the highly similar conformations are discarded in Musk1. In both data sets, a feature vector is explored for each conformation to illustrate its surface properties. During the learning procedure, a molecule is classified as a musk only if at least one conformation inside has a musky smell. Musk1 has 92 bags with a total of 476 instances. Of the bags, 47 are positive and 45 are negative. Musk2 contains 6,598 instances grouped into 102 bags, of which 39 are positive and 63 are negative. The instances in both data sets are described using a 166-dimensional feature vector.

4.3.2 Content-Based Image Classification Data

The content-based image classification MIL task determine whether an image contains an object of interest, e.g., an elephant or a tiger. This task is commonly used for MIL performance evaluation [29], [49], [50]. Our data are collected from the Corel data set [51]. All images in the set have been pre-processed and segmented by the Blobworld system [52]. Therefore, each image contains a set of small regions, which are described in terms of their color, texture, and shape characteristics (i.e., features). Each bag represents one image, and if one or more regions (instances) inside a bag contain the object of interest, the bag is labeled as positive, as shown in Fig. 7. In our experiments, each data set (Elephant or Tiger) consists of 100 positive and 100 negative bags, with a total of 1,391 and 1,220 instances, respectively. Each instance is represented by a 230-dimension feature vector.

4.3.3 Train Bound Challenge Data

The train bound challenge attempts to predict whether a train is eastbound or westbound. A train (bag) contains a variable number of cars (instances) of different shapes carrying different loads (features). Because a train's direction is either positive or negative under the MIL assumption, two train bound MIL data sets [18], [19], [53] are used for evaluation. The EastWest data uses 10 eastbound trains as the positive bags, while the WestEast data uses 10 westbound trains as

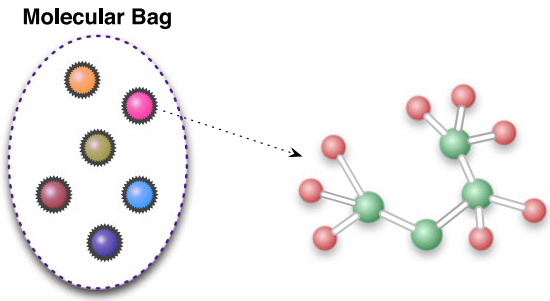


Fig. 6. Bag representation for the drug activity prediction task.

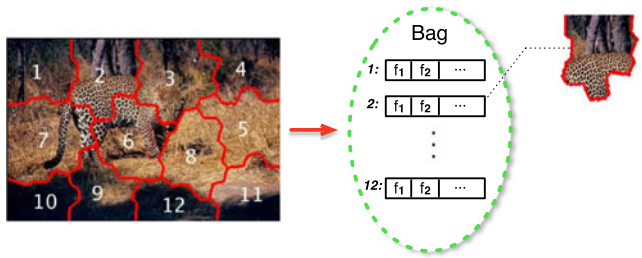


Fig. 7. Bag representation for the content-based image classification task.

positive bags. In other words, these two data sets have the same learning problem, but the class labels are reversed.

4.3.4 Mutagenicity Prediction Data

The mutagenesis data sets [54] describe a relational issue and have been widely used to explore inductive logic programming (ILP) tasks [55]. Mutagenicity predictions for a compound molecule are essentially predictions of carcinogenesis and, as such, the ability to identify these molecules is critical. Multiple instance learning frameworks have been successfully used to tackle this problem. In particular, by using the Proper toolbox [56], relational data can be represented as multiple instances in a bag by flattening the corresponding structure into an individual table. A bag represents one compound molecule in the mutagenesis MIL data sets, Atom and Bond, which contain all 1,618 atoms and all 3,995 atom-bond tuples as their instances, respectively. Each data set contains 125 positive and 63 negative bags.

4.3.5 Scientific Publication Retrieval Data

The DBLP data set consists of bibliographic data from the field of computer science.[1] Each record in DBLP, used in this experiment, is a paper published in the fields of either artificial intelligence (AI: IJCAI, AAAI, NIPS, UAI, COLT, ACL, KR, ICML, ECML, and IJCNN) or computer vision (CV: ICCV, CVPR, ECCV, ICPR, ICIP, ACM Multimedia, and ICME). The data set forms an MIL learning task with 100 positive (AI) and 100 negative bags [9]. A "bag-of-words" representation based on TFIDF [57] is used to convert the abstract of a paper into an instance with 4,497 features. Hence, each paper is a bag and each instance inside the bag denotes either the paper's abstract or the abstract of a reference cited in the paper. A conceptual view of constructing a multi-instance bag is shown in Fig. 8. The objective is to predict whether a paper belongs to the field of AI

1. http://dblp.uni-trier.de/xml/

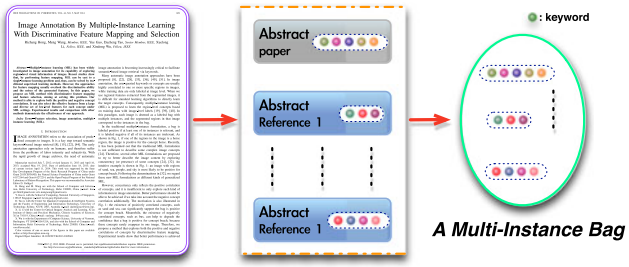


Fig. 8. Bag representation for the scientific publication retrieval task.

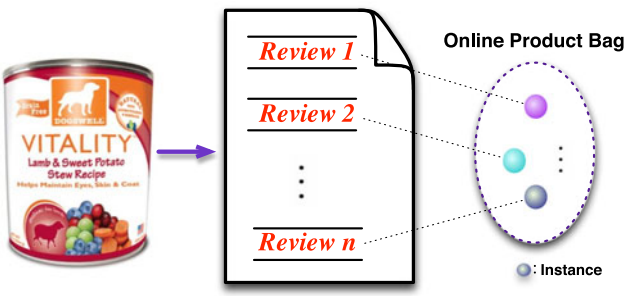


Fig. 9. Bag representation for the online product evaluation task.

or CV field using the abstract of each paper and the abstracts of its references. It is worth noting that the nature of AI and CV overlap in many aspects, such as machine learning, optimization, and visual information retrieval, which creates a challenging multi-instance learning task.

4.3.6 Online Product Evaluation Data

The online product evaluation task involves food reviews using the *Fine Foods* data from the Stanford Network Data Set Collection.[2] The data consists of numerous food related reviews from Amazon.com. Each review contains a product ID, a reviewer ID, a product score on a scale of 1 to 5, and detailed comments by the reviewer [58]. And each food product may have received multiple reviews. A product is considered interesting to other customers if one or more of its core characteristics, such as durability or affordability, has received an average review score ≥ 4 (very good). A score of < 4 across all reviews implies the product is not favored by customers. Our goal is to use the information in the review reports for online product evaluation. We choose 400 food products (2,097 reviews/instances with 1,517 features), each of which has received between 1 and 10 reviews, to form 200 positive bags (an average score ≥ 4) and 200 negative bags (every score < 4). An example of online product bag representation is shown in Fig. 9.

4.3.7 Newsgroup Categorization Data

The data set we choose uses the corpuses of two newsgroups as a base (rec.motorcycles and talk.politics.mideast).[3] Such types of data sets have been commonly used to evaluate the role of multiple instance learning frameworks [5], [29], [59]. In each news category, 3 percent of the posts are randomly selected from a target newsgroup category (e.g., rec.motorcycles) to generate the positive bags, along with posts drawn uniformly from other newsgroup categories. The articles are post-processed with stemming, stop-word removal, and information-gain ranked feature selection [60]. Then, the top 200 TFIDF [57] words are selected as features to represent each post (instance). Ultimately, each data set contains 100 bags with 50 being positive. The rec.motorcycles (shorten for News.rm) bag consists of 4,730 instances, and the talk.politics.mideast (shorten for News.tpm) bag contains 3,376 instances.

2. http://snap.stanford.edu/data/
3. http://people.csail.mit.edu/jrennie/20Newsgroups/

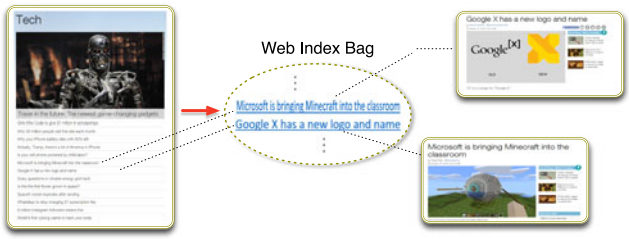


Fig. 10. Bag representation for the web index recommendation task.

TABLE 2
Compared Results in Terms of F-Measure and AUC with Their Standard Deviations for the Drug Activity Prediction Task

	F-measure		AUC	
	Musk1	Musk2	Musk1	Musk2
pMILGDM	0.857 ± 0.111	0.753 ± 0.105	0.859 ± 0.106	0.800 ± 0.108
aMILGDM	0.926 ± 0.090	0.795 ± 0.099	0.924 ± 0.092	0.834 ± 0.108
pMILLDM	0.935 ± 0.096	0.853 ± 0.085	0.935 ± 0.101	0.879 ± 0.110
aMILLDM	0.947 ± 0.065	0.857 ± 0.092	0.945 ± 0.070	0.890 ± 0.110
MILMR	0.854 ± 0.100	0.644 ± 0.103	0.835 ± 0.108	0.697 ± 0.109
MILWA	0.829 ± 0.102	0.696 ± 0.103	0.789 ± 0.109	0.744 ± 0.101
MILIR	0.748 ± 0.102	0.675 ± 0.104	0.703 ± 0.105	0.732 ± 0.103
MILES	0.863 ± 0.100	0.756 ± 0.102	0.846 ± 0.103	0.801 ± 0.103
pMILIS	0.805 ± 0.106	0.800 ± 0.109	0.817 ± 0.103	0.835 ± 0.106
aMILIS	0.826 ± 0.103	0.769 ± 0.103	0.826 ± 0.104	0.813 ± 0.102
MILFM	0.828 ± 0.095	0.773 ± 0.090	0.839 ± 0.098	0.816 ± 0.102

4.3.8 Web Index Recommendation Data

The web index recommendation task aims to recommend interesting web index pages to a particular user based on his or her preferences. Web pages from the Internet often have rich information, which is represented as a title or a brief summary with the details provided in linked pages. In practice, users may only indicate their interest in a page and not the specific content they prefer. For example, the technical web index page in Fig. 10 contains multiple concepts (e.g., cell phones, scholarships, traveling, and Google), but not all information on the page is likely to be of interest to the user. Such an observation naturally raises a multi-instance learning problem, where each web index page can be regarded as a bag with the linked pages inside as the instances. If one or more linked pages are of interest to the user, the page is considered positive, otherwise negative. The most frequent terms are explored as features to represent the instance/page.

4.4 Experimental Results

Tables 2-8 report the classification results in terms of F-measure and area under ROC curve with their standard deviations for the drug activity prediction task on Musk1 and Musk2, the content-based image classification task on Elephant and Tiger, the train bound challenge task on EastWest and WestEast, the mutagenicity prediction task on Atom

TABLE 3
Compared Results in Terms of F-Measure and AUC with Their Standard Deviations for the Content-Based Image Classification Task

	F-measure		AUC	
	Elephant	Tiger	Elephant	Tiger
pMILGDM	0.857 ± 0.066	0.770 ± 0.089	0.865 ± 0.074	0.770 ± 0.067
aMILGDM	0.845 ± 0.057	0.751 ± 0.092	0.855 ± 0.063	0.745 ± 0.072
pMILLDM	0.828 ± 0.070	0.730 ± 0.074	0.840 ± 0.082	0.760 ± 0.080
aMILLDM	0.843 ± 0.038	0.763 ± 0.102	0.845 ± 0.037	0.750 ± 0.097
MILMR	0.774 ± 0.104	0.714 ± 0.102	0.755 ± 0.091	0.700 ± 0.103
MILWA	0.709 ± 0.106	0.703 ± 0.109	0.590 ± 0.073	0.595 ± 0.089
MILIR	0.741 ± 0.101	0.609 ± 0.096	0.710 ± 0.098	0.615 ± 0.088
MILES	0.769 ± 0.089	0.720 ± 0.089	0.775 ± 0.089	0.720 ± 0.090
pMILIS	0.796 ± 0.102	0.589 ± 0.099	0.810 ± 0.099	0.665 ± 0.092
aMILIS	0.788 ± 0.101	0.737 ± 0.092	0.785 ± 0.098	0.750 ± 0.102
MILFM	0.757 ± 0.104	0.703 ± 0.084	0.785 ± 0.109	0.730 ± 0.094

TABLE 4
Compared Results in Terms of F-Measure and AUC with Their Standard Deviations for the Train Bound Challenge Task

	F-score		AUC	
	EastWest	WestEast	EastWest	WestEast
pMILGDM	0.857 ± 0.109	0.842 ± 0.109	0.850 ± 0.106	0.850 ± 0.106
aMILGDM	0.952 ± 0.106	0.947 ± 0.073	0.950 ± 0.026	0.950 ± 0.106
pMILLDM	0.800 ± 0.103	0.588 ± 0.107	0.800 ± 0.079	0.650 ± 0.108
aMILLDM	0.782 ± 0.100	0.667 ± 0.147	0.750 ± 0.117	0.700 ± 0.107
MILMR	0.667 ± 0.101	0.632 ± 0.115	0.650 ± 0.101	0.650 ± 0.132
MILWA	0.621 ± 0.100	0.400 ± 0.171	0.450 ± 0.139	0.250 ± 0.072
MILIR	0.500 ± 0.102	0.353 ± 0.126	0.500 ± 0.131	0.450 ± 0.144
MILES	0.600 ± 0.102	0.600 ± 0.130	0.600 ± 0.115	0.600 ± 0.115
pMILIS	0.696 ± 0.106	0.632 ± 0.126	0.650 ± 0.128	0.650 ± 0.149
aMILIS	0.526 ± 0.104	0.571 ± 0.134	0.550 ± 0.133	0.550 ± 0.145
MILFM	0.500 ± 0.103	0.625 ± 0.188	0.600 ± 0.122	0.700 ± 0.085

TABLE 5
Compared Results in Terms of F-Measure and AUC with Their Standard Deviations for the Mutagenicity Prediction Task

	F-measure		AUC	
	Atom	Bond	Atom	Bond
pMILGDM	0.891 ± 0.114	0.856 ± 0.102	0.861 ± 0.086	0.785 ± 0.104
aMILGDM	0.881 ± 0.109	0.847 ± 0.102	0.841 ± 0.082	0.757 ± 0.104
pMILLDM	0.881 ± 0.086	0.859 ± 0.111	0.812 ± 0.060	0.773 ± 0.103
aMILLDM	0.894 ± 0.081	0.861 ± 0.111	0.829 ± 0.063	0.789 ± 0.103
MILMR	0.825 ± 0.111	0.797 ± 0.103	0.733 ± 0.115	0.720 ± 0.103
MILWA	0.811 ± 0.117	0.812 ± 0.085	0.591 ± 0.222	0.599 ± 0.137
MILIR	0.820 ± 0.108	0.821 ± 0.093	0.749 ± 0.112	0.745 ± 0.101
MILES	0.879 ± 0.119	0.847 ± 0.095	0.801 ± 0.103	0.757 ± 0.102
pMILIS	0.829 ± 0.116	0.824 ± 0.101	0.761 ± 0.117	0.747 ± 0.103
aMILIS	0.827 ± 0.114	0.821 ± 0.100	0.765 ± 0.114	0.742 ± 0.105
MILFM	0.848 ± 0.114	0.853 ± 0.096	0.773 ± 0.112	0.709 ± 0.100

and Bond, the scientific publication retrieval task on AICV, the online product evaluation task on Food, the newsgroup categorization on News.rm and News.tpm, and the web index recommendation task on Web7 and Web8.

4.4.1 Comparisons with Non-Bag Mapping Instance Selection MIL Approaches

Among all the non-bag mapping instance selection methods, MILWA shows the worst performance in most cases. This is because MILWA assumes that all the instances in a bag share the bag's label. However, for a positive bag, at least one instance inside is positive, i.e., not all its instances must be positive. As a result, simply propagating the bag label to all the instances inside deteriorates classification performance. MILMR achieves a performance gain, mainly because of the mean strategy implemented on the bags. Nevertheless, for the drug activity prediction (e.g., Musk2 data) in Table 2, mutagenicity prediction (e.g., Bond data) in Table 5, newsgroup categorization (e.g., News.t.p.m data) in Table 7, and web index recommendation (e.g., Web8 data) in Table 8, the MILMR cannot achieve comparable classification performance to MILIR. This is partly because MILIR makes use of negative bag distribution. Overall, the experiments show that, of the instance selection methods, non-bag mapping methods do not perform as well as the bag mapping methods.

TABLE 6
Compared Results for the Scientific Publication Retrieval (AICV) and Online Product Evaluation (Food) Tasks

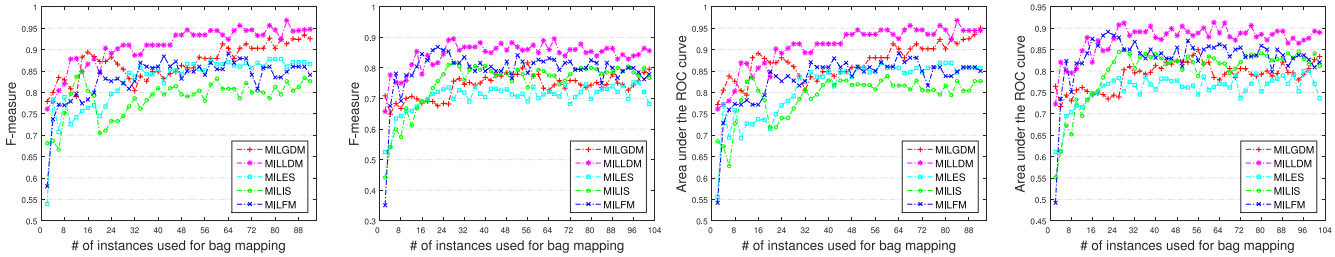
	F-measure		AUC	
	AICV	Food	AICV	Food
pMILGDM	0.764 ± 0.107	0.565 ± 0.101	0.720 ± 0.097	0.613 ± 0.087
aMILGDM	0.767 ± 0.105	0.582 ± 0.099	0.715 ± 0.078	0.648 ± 0.066
pMILLDM	0.806 ± 0.027	0.455 ± 0.037	0.820 ± 0.053	0.623 ± 0.049
aMILLDM	0.864 ± 0.107	0.535 ± 0.076	0.840 ± 0.090	0.623 ± 0.079
MILMR	0.436 ± 0.022	0.450 ± 0.069	0.565 ± 0.046	0.580 ± 0.075
MILWA	0.412 ± 0.109	0.420 ± 0.108	0.423 ± 0.093	0.565 ± 0.058
MILIR	0.445 ± 0.099	0.450 ± 0.087	0.555 ± 0.079	0.580 ± 0.081
MILES	0.667 ± 0.103	0.497 ± 0.052	0.655 ± 0.100	0.560 ± 0.035
pMILIS	0.699 ± 0.106	0.377 ± 0.079	0.725 ± 0.102	0.567 ± 0.060
aMILIS	0.760 ± 0.097	0.492 ± 0.064	0.765 ± 0.090	0.555 ± 0.079
MILFM	0.532 ± 0.024	0.375 ± 0.047	0.535 ± 0.034	0.567 ± 0.066

TABLE 7
Compared Results in Terms of F-Measure and AUC with Their Standard Deviations for the Newsgroup Categorization Task

	F-measure		AUC	
	News.rm	News.tpm	News.rm	News.tpm
pMILGDM	0.750 ± 0.103	0.706 ± 0.101	0.780 ± 0.107	0.750 ± 0.108
aMILGDM	0.767 ± 0.115	0.733 ± 0.101	0.800 ± 0.101	0.760 ± 0.108
pMILLDM	0.764 ± 0.096	0.758 ± 0.101	0.790 ± 0.105	0.790 ± 0.106
aMILLDM	0.651 ± 0.113	0.713 ± 0.097	0.700 ± 0.105	0.750 ± 0.107
MILMR	0.427 ± 0.095	0.457 ± 0.074	0.570 ± 0.107	0.620 ± 0.102
MILWA	0.447 ± 0.098	0.490 ± 0.091	0.505 ± 0.102	0.505 ± 0.101
MILIR	0.538 ± 0.088	0.571 ± 0.083	0.640 ± 0.099	0.670 ± 0.103
MILES	0.650 ± 0.116	0.635 ± 0.101	0.710 ± 0.103	0.690 ± 0.108
pMILIS	0.643 ± 0.117	0.628 ± 0.127	0.700 ± 0.106	0.680 ± 0.106
aMILIS	0.643 ± 0.114	0.642 ± 0.098	0.700 ± 0.107	0.690 ± 0.108
MILFM	0.651 ± 0.101	0.659 ± 0.106	0.700 ± 0.101	0.710 ± 0.103

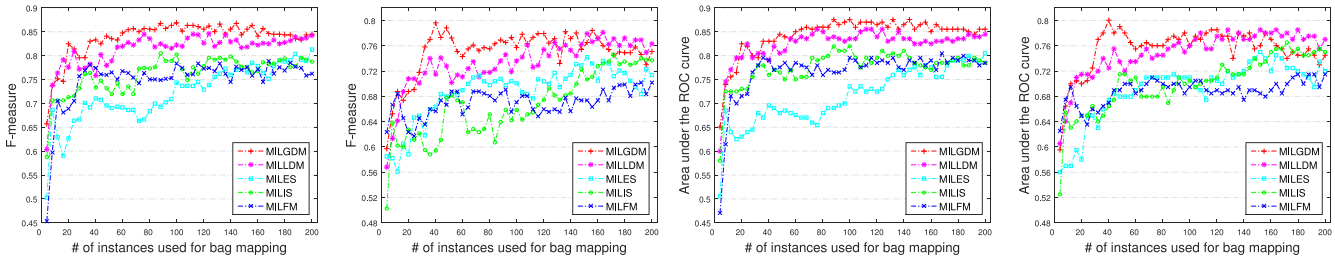
TABLE 8
Compared Results in Terms of F-Measure and AUC with Their Standard Deviations for the Web Index Recommendation Task

	F-measure		AUC	
	Web7	Web8	Web7	Web8
pMILGDM	0.667 ± 0.109	0.660 ± 0.106	0.704 ± 0.105	0.680 ± 0.101
aMILGDM	0.660 ± 0.106	0.634 ± 0.108	0.695 ± 0.105	0.670 ± 0.105
pMILLDM	0.689 ± 0.069	0.703 ± 0.095	0.752 ± 0.103	0.756 ± 0.107
aMILLDM	0.723 ± 0.106	0.707 ± 0.108	0.764 ± 0.102	0.741 ± 0.102
MILMR	0.661 ± 0.102	0.647 ± 0.102	0.595 ± 0.107	0.564 ± 0.091
MILWA	0.662 ± 0.102	0.654 ± 0.104	0.534 ± 0.067	0.508 ± 0.081
MILIR	0.649 ± 0.097	0.468 ± 0.101	0.570 ± 0.087	0.629 ± 0.082
MILES	0.595 ± 0.101	0.571 ± 0.103	0.602 ± 0.108	0.600 ± 0.105
pMILIS	0.518 ± 0.086	0.512 ± 0.105	0.627 ± 0.066	0.639 ± 0.100
aMILIS	0.667 ± 0.102	0.661 ± 0.102	0.665 ± 0.103	0.656 ± 0.109
MILFM	0.505 ± 0.094	0.416 ± 0.081	0.597 ± 0.103	0.596 ± 0.102



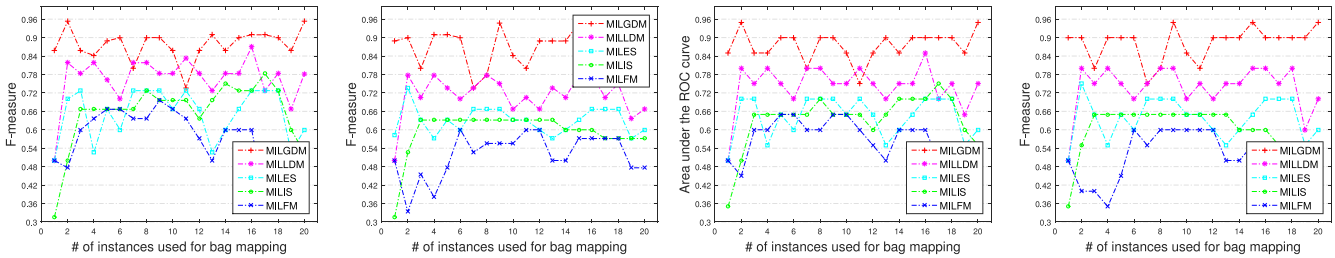
(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

Fig. 11. Bag mapping performance comparisons with different sizes of IIP or DIP for drug activity prediction: (a) Musk1 and (b) Musk2.



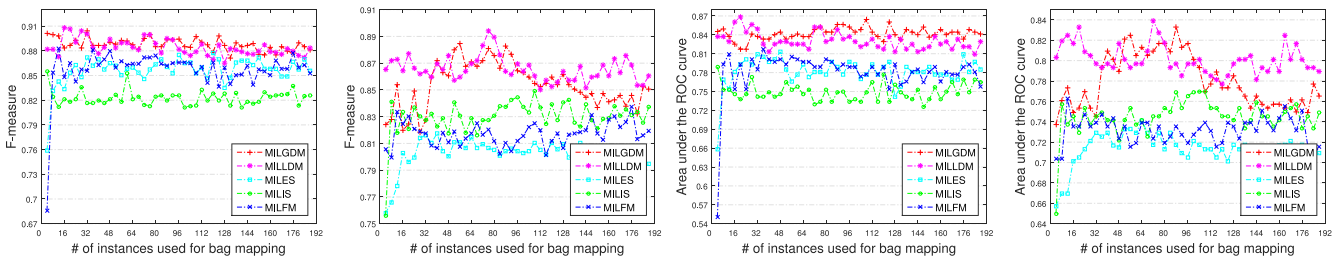
(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

Fig. 12. Bag mapping performance comparisons with different sizes of IIP or DIP for content-based image classification: (a) Elephant and (b) Tiger.



(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

Fig. 13. Bag mapping performance comparisons with different sizes of IIP or DIP for train bound challenge: (a) EastWest and (b)WestEast.



(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

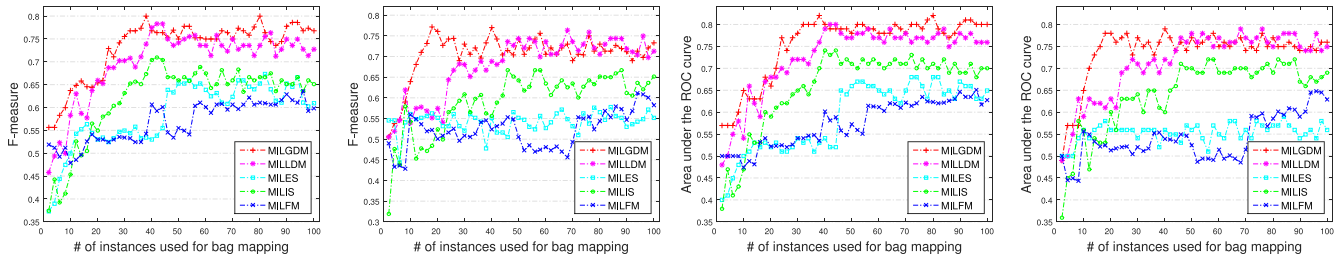
Fig. 14. Bag mapping performance comparisons with different sizes of IIP or DIP for mutagenicity prediction: (a) Atom and (b) Bond.

4.4.2 *Comparisons with Bag Mapping Instance Selection MIL Approaches*

Within the bag mapping instance selection methods, MILES often outperforms MILFM, especially on the scientific publication retrieval task. As shown in Table 6, MILES achieves a 66.7 percent F-measure and a 65.5 percent AUC—much higher than MILFM's F-measure (53.2 percent) and AUC (53.3 percent)—with similar observations for the other learning tasks, like mutagenicity prediction (e.g., Atom) in Table 5 and web index recommendation in Table 8. Such superiority of MILES is mainly attributed to full use of the instances for bag mapping. However, MILES cannot achieve better classification performance than MILFM in some cases, for instance, on the train bound challenge WestEast data in Table 4. This suggests that not all instances contribute to the final classification performance. Although MILFM uses the clusters in negative bags to overcome this issue, it still uses all the instances in the positive bags without further improvement. Accordingly, MILIS (pMILIS or aMILIS), which uses an instance pruning strategy based on kernel density estimation, performs better than the other two bag mapping approaches, i.e., MILES with non-instance pruning and MILFM with partial instance pruning.

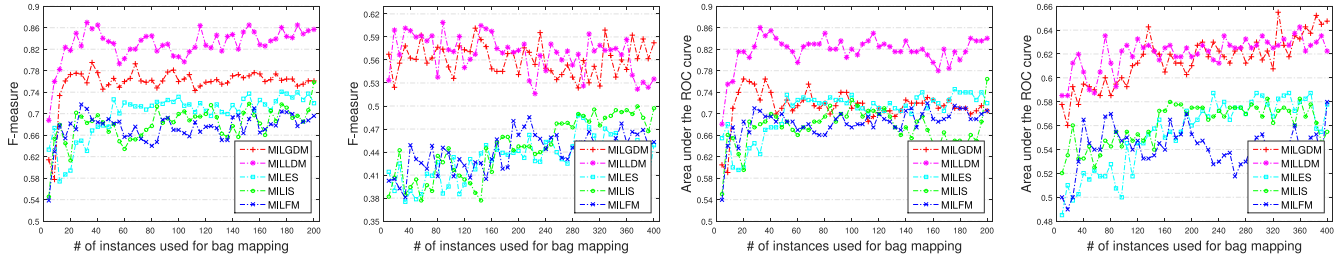
4.4.3 *Comparisons with Discriminative Bag Mapping MIL Approaches*

Compared to MIL methods with instance selection, the four approaches based on discriminative bag mapping



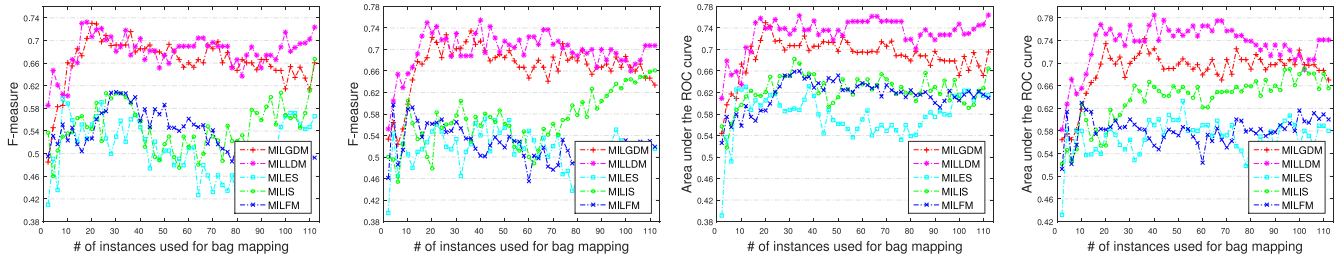
(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

Fig. 15. Bag mapping performance comparisons with different sizes of IIP or DIP for newsgroup categorization: (a) News.rm and (b) News.tpm.



(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

Fig. 16. Bag mapping performance comparisons with different sizes of IIP or DIP for (a) scientific publication retrieval and (b) online product evaluation.



(a).1 F-measure (b).1 F-measure (a).2 AUC (b).2 AUC

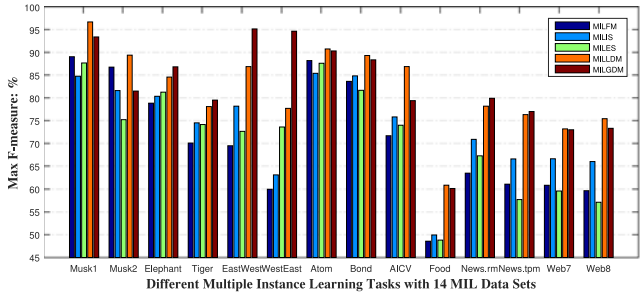
Fig. 17. Bag mapping performance comparisons with different sizes of IIP or DIP for web index recommendation: (a) Web7 and (b) Web8.

demonstrate better performance on all data sets. For example, the proposed local discriminative bag mapping aMILLDM, which is based on all bags, achieves the highest classification performance for drug activity prediction (e.g., Musk1) in Table 2, mutagenicity prediction (e.g., Bond) in Table 5, and web index recommendation (e.g., Web7) in Table 8. The proposed global discriminative bag mapping aMILGDM based on all bags and pMILGDM based on only positive bags achieve the highest classification performance for content-based image classification in Table 3 and the train bound challenge in Table 4. aMILGDM and aMILLDM's performance is comparable. In summary, aMILDM (aMILGDM or aMILLDM), which constructs the DIP using all bags, outperforms pMILDM (pMILGDM or pMILLDM) using only positive bags, because more information is used to construct the DIP.

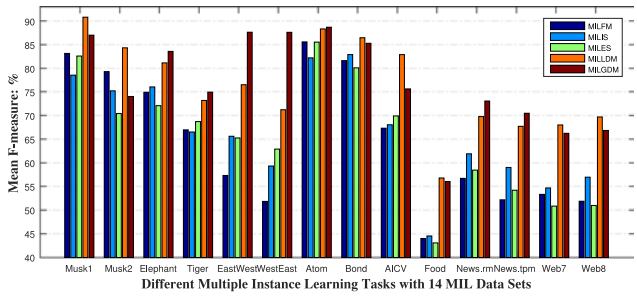
4.4.4 *Bag Mapping Performance Comparisons* w.r.t. *Different Size of IIP or DIP*

In terms of the transitional bag mapping MIL methods, MILIS with different sizes of IIP (aMILIS with positive bags or pMILIS with all bags) achieves a range of classification performance results with both the global or local discriminative bag mapping MIL methods. Figs. 11-17 report the bag mapping classification performance with respect to different sizes of IIP or DIP. The number of instances used in bag mapping is varied from one to the number of bags provided for each data on the eight different types of learning tasks. As the number of instances increases, the classification performance improves. That is because the new instances provide further information that is useful to the bag mapping. When the instances in the IIP or DIP are not adequate, the rising trend in performance is insignificant. The train bound challenge data with only 20 bags is a good example of this. MILES and MILFM achieve comparable F-measure and AUC scores but are inferior to MILIS, which uses only one instance from each bag for bag mapping (i.e., instance pruning). For instance, when the size of the IIP is greater than 20, MILIS continuously outperforms MILES and MILFM. However, all the three types of bag mapping approaches, including the state-of-the-art MILFM, cannot match the performance of the proposed discriminative bag mapping method in either of its local (MILLDM) or global (MILGDM) variants.

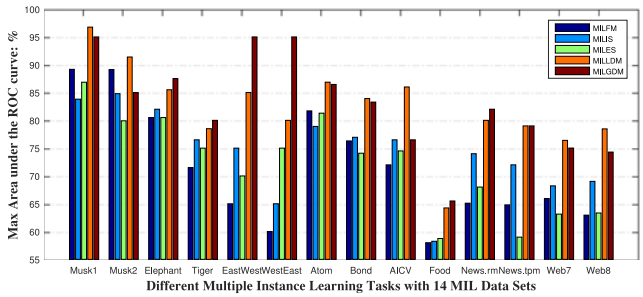
Fig. 18 reports the maximum and average classification performance for the local and global DIP-based discriminative bag mapping MILDM compared to traditional IIP bag mapping approaches. Figs. 18a.1 and 18a.2 show the max and average F-measure for a range of IIP/DIP sizes. Figs. 18b.1 and 18b.2 show the same in terms of AUC. According to the results, our proposed MILLDM and MILGDM show more improvement over the other three



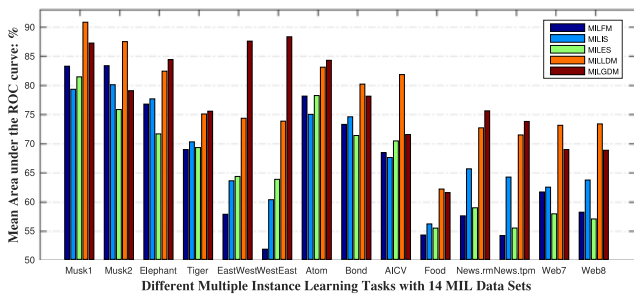
(a).1 F-measure



(a).2 F-measure



(b).1 AUC



(b).2 AUC

Fig. 18. Maximum and average classification performance (F-measure and AUC) for the local and global DIP-based discriminative bag mapping methods (MILLDM and MILGDM) versus traditional IIP bag mapping approaches, MILES, MILIS, and MILFM, for eight different types of multiple instance learning tasks across 14 MIL data sets.

TABLE 9
Pairwise t-Test Results for MILDM (Global MILGDM or Local MILLDM) versus General Bag Mapping MIL Methods on Eight Types of Learning Tasks across 14 Data Sets

(a) t-test on F-measure

Data Sets	A1-A2	A1-B	A1-C	A1-D	A2-B	A2-C	A2-D
Musk1	6.43e-09	4.89e-07	2.75e-17	8.11e-07	1.56e-20	5.35e-27	4.33e-21
Musk2	6.91e-25	2.76e-07	1.80e-01	1.12e-05	8.28e-35	2.86e-19	1.36e-08
Elephant	1.07e-17	3.13e-44	2.54e-48	3.18e-53	1.25e-39	6.25e-41	3.59e-38
Tiger	8.18e-08	1.10e-35	4.87e-31	5.93e-41	4.78e-30	1.33e-35	1.16e-32
EastWest	4.18e-05	3.09e-09	1.25e-07	4.95e-10	5.14e-07	1.19e-05	2.80e-08
WestEast	9.63e-08	1.59e-11	1.25e-10	1.12e-12	3.60e-06	1.05e-08	6.46e-08
Atom	1.55e-02	4.88e-25	4.99e-66	1.24e-13	4.24e-23	4.11e-61	1.51e-13
Bond	3.65e-05	1.58e-26	3.66e-11	1.29e-15	1.07e-31	2.64e-18	4.27e-30
AICV	1.35e-24	1.11e-11	3.64e-20	1.51e-26	2.79e-27	8.59e-40	5.52e-43
Food	1.13e-01	8.87e-29	3.24e-22	9.64e-31	4.44e-28	9.38e-22	2.06e-28
News.rm	2.33e-10	4.69e-28	3.87e-29	4.11e-29	5.34e-22	1.29e-31	1.06e-21
News.tpm	3.67e-03	8.73e-27	1.34e-17	5.92e-28	1.25e-17	3.45e-25	5.57e-20
Web7	3.02e-04	4.68e-27	4.69e-23	3.70e-29	5.52e-32	2.75e-30	1.50e-31
Web8	9.83e-09	1.21e-31	3.94e-18	3.25e-29	1.02e-40	2.97e-22	9.61e-35

(b) t-test on AUC

Data Sets	A1-A2	A1-B	A1-C	A1-D	A2-B	A2-C	A2-D
Musk1	1.66e-07	2.27e-09	7.61e-15	1.44e-06	8.41e-23	6.46e-27	3.71e-19
Musk2	3.93e-25	6.44e-08	1.60e-01	6.42e-06	9.19e-35	3.07e-19	3.98e-09
Elephant	8.58e-17	1.35e-48	7.75e-54	5.51e-49	1.04e-42	2.39e-45	5.13e-36
Tiger	8.74e-02	4.40e-31	1.67e-24	8.27e-38	1.90e-42	1.56e-35	1.62e-37
EastWest	1.49e-06	6.55e-11	1.89e-09	1.88e-12	1.81e-07	2.67e-06	8.92e-09
WestEast	1.83e-06	5.84e-10	2.60e-10	1.59e-12	1.81e-07	1.05e-08	1.96e-08
Atom	1.70e-05	5.12e-36	9.87e-62	2.41e-22	1.34e-34	3.62e-61	7.35e-19
Bond	2.04e-05	1.49e-24	6.82e-12	2.75e-14	8.94e-30	4.44e-20	1.11e-29
AICV	1.87e-30	1.38e-01	9.08e-09	2.01e-10	6.56e-24	5.13e-38	4.26e-39
Food	8.07e-03	8.38e-28	4.98e-27	7.45e-27	3.24e-26	1.73e-33	1.55e-33
News.rm	6.51e-09	1.75e-26	4.22e-26	4.41e-30	3.13e-24	5.58e-28	1.45e-25
News.tpm	3.48e-03	8.79e-30	7.35e-16	8.39e-27	1.37e-21	1.26e-26	1.19e-22
Web7	5.73e-17	7.82e-27	2.20e-25	3.05e-26	8.41e-34	7.42e-37	5.42e-40
Web8	1.53e-17	5.06e-29	3.12e-16	6.98e-27	5.59e-38	2.06e-23	1.11e-31

A1 and A2 denote the proposed DIP-based global and local discriminative bag mapping MILGDM and MILLDM, respectively. B, C, and D denote IIP-based MILES, MILIS, and MILFM, respectively.

TABLE 10
Time Complexity: Bag Mapping Instance Selection Approaches

pMILDM	aMILDM	MILES	pMILIS	aMILIS	MILFM
$O(p^+n^2+n^+pd)$	$O(pn^2+npd)$	$O(p^2d)$	$O(p^-p^+d+n^+pd)$	$O(p^-pd+npd)$	$O(p^-n^+t+p^+d+n^+d)$

TABLE 11
Average CPU Running Time for the Compared Algorithms in the Training Phase on Eight MIL Learning Tasks (Measured in Milliseconds)

	Musk2	Elephant	EastWest	Bond	AICV	Food	News.rm	Web7
	# of Ins: 6,598 # of Att: 166	# of Ins: 1,391 # of Att: 230	# of Ins :213 # of Att: 24	# of Ins :3,995 # of Att:16	# of Ins :1,151 # of Att: 4,497	# of Ins :2,097 # of Att: 1,517	# of Ins :4,730 # of Att: 200	# of Ins :15,000 # of Att: 26
pMILGDM	3,241	1,015	72	1,171	8,424	11,080	6,540	99,467
aMILGDM	19,930	1,726	82	1,544	16,348	20,058	12,382	201,491
pMILLDM	3,302	1,016	71	1,142	8,433	11,236	6,474	99,459
aMILLDM	19,815	1,712	76	1,502	16,513	20,089	12,393	200,234
MILMR	314	534	92	175	43,621	13,291	359	71,175
MILWA	77,605	5,563	142	3,049	30,887	115,922	45,185	2,183,526
MILIR	13,988	714	87	414	8,775	9,547	5,117	93,861
MILES	17,169	1,838	148	2,752	12,411	17,405	10,488	186,948
pMILIS	2,441	613	95	501	4,442	5,140	2,775	48,910
aMILIS	14,185	936	114	633	8,522	9,465	5,322	98,410
MILFM	35,070	4,964	185	2,611	69,889	49,227	16,308	492,891
MISVM	121,326	19,070	278	24,613	110,461	215,016	21,225	7,271,408
MILR	80,693	15,686	108	1,542	5,513	125,503	2,530	2,629,114
MIEMDD	1,182,981	667,648	1,324	12,730	134,413	3,439,775	217,528	1,511,795
MIBoost	21,641	4,509	101	1,272	239,730	68,150	11,470	1,890,149

baselines, and a comparable result between the two. On average, MILDM is 5-25 percent more accurate for classification performance in terms of F-measure than the traditional IIP bag mapping methods (e.g., around 20 percent improvements on the train bound challenge task and the online product evaluation task, and 10 percent improvements on the newsgroup categorization task and the web index recommendation task as shown in Fig. 18a). Similar observations can be found in terms of AUC classification performance in Fig. 18b. This demonstrates that, by using the DIP evaluation criteria, MILDM is able to find the most effective instances to map the instance bags for classification.

4.4.5 Statistical Significance Comparisons

In reality, however, MILDM may not always achieve good performance. For example, the global MILGDM does not perform as well as MILFM when the size of the IIP/DIP drops to between 20 and 24 (Figs. 11b.1 and 11b.2). To confirm that this observation has no effect on the superiority of the discriminative bag mapping approaches, we conduct t-test results on the MILDM in Table 9, and summarize a two-tailed t-test between the MILDM and the traditional bag mapping approaches, MILES, MILIS, and MILFM. All the pairwise t-test values are calculated with a 95 percent level of confidence ($\alpha = 0.05$). Each value in Table 9 is the p-value for a pairwise t-test between two learning algorithms. According to statistical theory, the proposed MILDM has achieved a statistically significant improvement compared to the other bag mapping methods if the p-value is less than 0.05.

From the second column in each subtable (e.g., Table 9a), the difference between MILLDM and MILGDM is statistically significant with most of the p-values at less than 0.05. The exceptions are their performance on the online product evaluation task in terms of F-measure, and the content-based image classification task (e.g., Tiger) in terms of AUC performance where their performance is comparable. Compared to the traditional bag mapping approaches, MILDM outperforms MILES, MILIS, and the state-of-the-art MILFS in all cases.

4.4.6 Time Complexity Analysis

The main components of the MILDM's time complexity include constructing the DIP and bag mapping. Constructing the DIP for aMILDM costs $O(pn^2)$, where p represents the number of instances in all bags. The bag mapping procedure costs $O(npd)$, where d denotes the dimensions of the data. By contrast, pMILDM has a complexity of $O(p^+n^2+n^+pd)$, with p^+ denoting the number of instances in the positive bags and n^+ denoting the size of positive bags. Table 10 summarizes the time complexities of the other bag mapping approaches with instance selection. MILES uses all the instances as the IIP for bag mapping with $O(p^2d)$ computational complexity. pMILIS requires a kernel density estimation based on the distribution of the negative instances to select one instance from the positive bag for the IIP with $O(p^-p^+d)$ complexity, where p^- denotes the number of instances in the negative bags. Together with the bag mapping complexity $O(n^+pd)$, pMILIS's total time complexity is $O(p^-p^+d+n^+pd)$. Similarly, aMILIS costs $O(p^-pd+npd)$. MILFM uses the instances in positive bags and the clusters of negative instances, which costs $O(p^-n^+t+p^+d+n^+d)$, where the t denotes the number of iterations during the clustering, $O(p^-n^+t)$ represents the clustering process, and $O(p^+d+n^+d)$ represents the bag mapping.

4.4.7 Efficiency Comparisons

Table 11 reports the average CPU runtime performance of the training phase on the eight different multi-instance

TABLE 12
Average CPU Running Time for the Compared Algorithms in the Testing Phase on Eight MIL Learning Tasks (Measured in Milliseconds)

	Musk2	Elephant	EastWest	Bond	AICV	Food	News.rm	Web7
	# of Ins: 6,598 # of Att: 166	# of Ins: 1,391 # of Att: 230	# of Ins :213 # of Att: 24	# of Ins :3,995 # of Att:16	# of Ins :1,151 # of Att: 4,497	# of Ins :2,097 # of Att: 1,517	# of Ins :4,730 # of Att: 200	# of Ins :15,000 # of Att: 26
pMILGDM pMILLDM pMILIS	205	282	60	242	1,305	1,461	216	2,002
aMILGDM aMILLDM aMILIS	358	394	80	295	1,730	2,868	321	3,713
MILMR	113	232	42	173	1,846	2,349	125	7,985
MILWA	3,605	482	75	486	1,961	4,287	812	31,547
MILIR	1,554	252	87	200	1,135	1,936	653	10,507
MILES	2,037	377	47	460	1,539	2,809	1,250	20,850
MILFM	852	189	29	345	792	1,568	657	10,925
MISVM	508	112	14	117	298	628	280	17,525
MILR	96	115	18	108	84	31	15	143
MIEMDD	86	79	72	75	141	117	84	300
MIBoost	129	148	15	128	2,314	511	262	17,463

learning tasks. Each data set, e.g., Food, represents a specific MIL learning task, e.g., online product evaluation. When the number of instances or attributes is small, as in the EastWest data with 24 attributes or the Bond data with 16 attributes, the runtime between the proposed pMILDM with only positive bags and aMILDM with all bags is similar. However, as the number of instances or attributes increases, pMILDM achieves a much better runtime performance than aMILDM. This is understandable given pMILDM uses fewer bags to construct the DIP.

The runtime performance of non-bag mapping approaches slightly outperforms the bag mapping approaches. This is largely because the bag mapping processing requires extra learning time. However, the proposed MILDM takes less runtime than the non-bag mapping with a large amount of data. For example, pMILDM's runtime on scientific publication retrieval AICV data is around 8,000 milliseconds, whereas MILMR and MILWA takes four time as long (about 32,000 milliseconds). Among all bag mapping methods, the state-of-the-art MILFM has the worse runtime performance, mainly attributed to the time-consuming clustering procedures on the negative bags. Compared to the traditional bag mapping approach MILIS, pMILDM achieves comparable runtime performance. Furthermore, pMILDM significantly outperforms MILES for runtime efficiency, because MILES uses all the instances in the bags to build an IIP for mapping. By contrast, pMILDM only selects one instance from each positive bag to build the DIP. In summary, the discriminative bag mapping MILDM achieves a good balance between runtime efficiency and classification effectiveness.

Table 12 reports the average CPU runtime performance for the testing phase. After the DIP or IIP is constructed in the training phase, the instance selection bag mapping methods directly use the DIP or IIP for test data bag mapping. In this case, the instance selection-based methods with the same sized DIP or IIP (e.g., pMILGDM, pMILLDM and pMILIS) have the same testing time. In other words, among the bag mapping instance selection-based algorithms, the corresponding testing time depends on m (the size of DIP or IIP). The larger the m, the more testing time the algorithm needs. For instance, MILES has the worse runtime performance among the bag mapping methods because it uses all the instances for mapping.

5 Discussion

5.1 MILDM with Different Base Classifiers

To demonstrate that MILDM is effective for different learning algorithms, four representative classifiers are used for validation: k-nearest neighbors (IB1), naive Bayes (NB), decision trees (J48), and support vector machines (SMO). Table 13 reports the maximum AUC achieved by these four versions of MILDM. The results show that MILDM using IB1 achieves the highest performance, the alternative base classifiers achieve more or less comparable classification performance on all three data sets. However, none of the alternatives produces consistently better performance on the three data sets than the MILDM with IB1.

5.2 Comparisons to MIL without Instance Selection

Our experiments show that MILDM achieves the best performance of all the instance selection MIL methods. In this section, we report the performance of MILDM compared to four MIL algorithms—MISVM [49], MILR [27], MIEMDD [31], and MIBoost [28] (Table 14). MISVM is an

TABLE 13
Performance Comparison (AUC) When Using Different Supervised Learning Algorithms as the Base Classifiers in MILDM

	Elephant	Bond	AICV
MILDM+IB1	0.865 ± 0.074	0.789 ± 0.103	0.840 ± 0.090
MILDM+NB	0.783 ± 0.071	0.713 ± 0.059	0.782 ± 0.101
MILDM+SMO	0.820 ± 0.094	0.754 ± 0.082	0.806 ± 0.032
MILDM+J48	0.814 ± 0.115	0.764 ± 0.135	0.775 ± 0.108

TABLE 14
Compared Results of MILDM (Global MILGDM or Local MILLDM) versus General MIL Methods on Eight Types of Learning Tasks across 14 Data Sets

(a) F-measure

Data Sets	MILDM	MISVM	MILR	MIEMDD	MIBoost
Musk1	0.947 ± 0.065	0.716 ± 0.192	0.742 ± 0.215	0.855 ± 0.143	0.839 ± 0.132
Musk2	0.857 ± 0.092	0.785 ± 0.117	0.803 ± 0.047	0.817 ± 0.102	0.789 ± 0.115
Elephant	0.857 ± 0.066	0.794 ± 0.086	0.749 ± 0.151	0.728 ± 0.090	0.743 ± 0.142
Tiger	0.770 ± 0.089	0.724 ± 0.163	0.745 ± 0.057	0.709 ± 0.134	0.716 ± 0.104
EastWest	0.952 ± 0.106	0.714 ± 0.122	0.667 ± 0.477	0.794 ± 0.085	0.726 ± 0.038
WestEast	0.947 ± 0.073	0.786 ± 0.211	0.545 ± 0.483	0.765 ± 0.066	0.743 ± 0.012
Atom	0.894 ± 0.081	0.799 ± 0.180	0.797 ± 0.187	0.808 ± 0.051	0.753 ± 0.133
Bond	0.861 ± 0.111	0.782 ± 0.127	0.820 ± 0.183	0.804 ± 0.073	0.736 ± 0.147
AICV	0.864 ± 0.107	0.808 ± 0.069	0.819 ± 0.143	0.767 ± 0.068	0.730 ± 0.062
Food	0.582 ± 0.099	0.545 ± 0.063	0.544 ± 0.108	0.562 ± 0.101	0.561 ± 0.075
News.rm	0.767 ± 0.115	0.765 ± 0.131	0.714 ± 0.106	0.729 ± 0.038	0.744 ± 0.066
News.tpm	0.758 ± 0.101	0.726 ± 0.141	0.713 ± 0.212	0.701 ± 0.084	0.745 ± 0.072
Web7	0.723 ± 0.106	0.701 ± 0.172	0.675 ± 0.227	0.686 ± 0.028	0.619 ± 0.134
Web8	0.707 ± 0.108	0.697 ± 0.136	0.632 ± 0.156	0.677 ± 0.076	0.629 ± 0.132

(b) AUC

Data Sets	MILDM	MISVM	MILR	MIEMDD	MIBoost
Musk1	0.945 ± 0.070	0.759 ± 0.135	0.751 ± 0.230	0.905 ± 0.092	0.838 ± 0.109
Musk2	0.890 ± 0.110	0.730 ± 0.142	0.862 ± 0.080	0.846 ± 0.071	0.859 ± 0.100
Elephant	0.865 ± 0.074	0.780 ± 0.086	0.831 ± 0.111	0.810 ± 0.089	0.772 ± 0.122
Tiger	0.770 ± 0.067	0.740 ± 0.123	0.750 ± 0.062	0.727 ± 0.041	0.721 ± 0.114
EastWest	0.950 ± 0.026	0.760 ± 0.211	0.710 ± 0.222	0.769 ± 0.098	0.785 ± 0.047
WestEast	0.950 ± 0.106	0.754 ± 0.258	0.400 ± 0.116	0.749 ± 0.048	0.751 ± 0.012
Atom	0.861 ± 0.086	0.772 ± 0.145	0.776 ± 0.121	0.759 ± 0.062	0.858 ± 0.095
Bond	0.789 ± 0.103	0.751 ± 0.089	0.818 ± 0.073	0.727 ± 0.033	0.750 ± 0.085
AICV	0.840 ± 0.090	0.785 ± 0.123	0.806 ± 0.079	0.794 ± 0.098	0.809 ± 0.100
Food	0.648 ± 0.066	0.615 ± 0.073	0.603 ± 0.126	0.612 ± 0.088	0.607 ± 0.110
News.rm	0.800 ± 0.101	0.770 ± 0.125	0.709 ± 0.087	0.730 ± 0.071	0.710 ± 0.009
News.tpm	0.790 ± 0.106	0.710 ± 0.110	0.732 ± 0.154	0.752 ± 0.086	0.726 ± 0.063
Web7	0.764 ± 0.102	0.692 ± 0.173	0.699 ± 0.169	0.660 ± 0.026	0.733 ± 0.149
Web8	0.756 ± 0.107	0.686 ± 0.097	0.619 ± 0.125	0.610 ± 0.107	0.726 ± 0.103

MILDM shows the best performance achieved by the global MILGDM or local MILLDM.

TABLE 15
Compared Results on Large-Scale Speaker Data in Terms of F-Measure and AUC

	pMILGDM	aMILGDM	pMILLDM	aMILLDM	MILMR	MILWA	MILIR	MILES	pMILIS	aMILIS	MILFM
F-measure	0.961	0.985	0.894	0.932	0.750	0.724	0.800	0.900	0.864	0.889	0.822
AUC	0.952	0.972	0.883	0.917	0.778	0.738	0.747	0.909	0.812	0.858	0.830

implementation of support vector machines [61] for MIL, MILR is a logistics-based learning method, MIEMDD is an improved diverse density approach, and MIBoost is an algorithm inspired by AdaBoost. Although MISVM, MILR, MIEMDD, and MIBoost achieve good performance, they still cannot reach the best performance of the proposed discriminative bag mapping MILDM. In addition, the CPU runtime comparisons on training and test phrases are respectively reported in Tables 11 and 12. The results confirm that the proposed MILDM requires less runtime than the MIL without instance selection on a large amount of data.

5.3 Scalability of MILDM

We compare the proposed methods with the non-bag mapping instance selection approaches (MILMR, MILWA and MILIR) and the bag mapping instance selection approaches (MILES, pMILIS, aMILIS and MILFM) on the large-scale "Speaker" data set which includes 430 bags with 583,600 instances. The numbers of instances per bag are 1,357. More details are available in [62]. With this data, 70 percent of the data is used for training with the remainder for testing. Table 15 shows the F-measure and AUC classification performance results. We can see that the proposed MILDM approach still exhibits clear advantages in terms of these two evaluation metrics.

6 Conclusion

This paper investigates an instance selection-based multiple-instance bag mapping task, where each bag is mapped to a new feature space using a small number of selected instances from multi-instance bags. The mapped instances can be directly used by generic learning algorithms to train classifiers for solving multiple-instance learning tasks. Due to multi-instance bag constraints, determining good instances for bag mapping is difficult. To this end, we proposed a discriminative bag mapping approach that builds a discriminative instance pool to ensure the bags in the new mapping

space can be easily separated from each other. Experiments and comparisons on eight types of real-world multiple-instance learning tasks (including 14 data sets) demonstrate a consistent performance gain. The proposed MILDM outperforms the state-of-the-art MIL bag mapping approaches in terms of F-measure and AUC. A CPU runtime performance study further demonstrates that MILDM provides an effective trade-off between runtime efficiency and classification effectiveness.

Acknowledgments

The work was supported by the US National Science Foundation (NSF) through Grant IIS-1652107, the Australian Research Council (ARC) through Grants DP140102206 and DP140100545, the MQNS Grant No. 9201701203, and the MQ Enterprise Partnership Scheme Pilot Res Grant No. 9201701455.

References

[1] T. Dietterich., R. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, pp. 31–71, 1997.
[2] R. Hong, W. Meng, G. Yue, D. Tao, X. Li, and X. Wu, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.
[3] Y. Chen, J. Bi, and J. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.
[4] H. Yuan, M. Fang, and X. Zhu, "Hierarchical sampling for multi-instance ensemble learning," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2900–2905, Dec. 2013.
[5] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," in *Proc. 29th Int. Conf. Mach. Learn.*, 2009, pp. 1249–1256.
[6] D. Kelly, J. McDonald, and C. Markham, "Weakly supervised training of a sign language recognition system using multiple instance learning density matrices," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 41, no. 2, pp. 526–541, Apr. 2011.
[7] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.
[8] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence, "Multiple instance learning on structured data," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 145–153.
[9] J. Wu, X. Zhu, C. Zhang, and P. Yu, "Bag constrained structure pattern mining for multi-graph classification," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2382–2396, Oct. 2014.
[10] J. Wu, S. Pan, X. Zhu, and Z. Cai, "Boosting for multi-graph classification," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 430–443, Mar. 2015.
[11] Z.-H. Zhou, K. Jiang, and M. Li, "Multi-instance learning based web mining," *Appl. Intell.*, vol. 22, no. 2, pp. 135–147, 2005.
[12] Z. Zhou, "Multi-instance learning: A survey," Nanjing University, Nanjing, China, 2004.
[13] J. R. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowl. Eng. Rev.*, vol. 25, no. 1, pp. 1–25, 2010.
[14] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, 2013.
[15] E. Alpaydin, V. Cheplygina, M. Loog, and D. M. Tax, "Single- versus multiple-instance classification," *Pattern Recogn.*, vol. 48, no. 9, pp. 2831–2838, 2015.
[16] Z. Fu, A. Robles-Kelly, and J. Zhou, "MILIS: Multiple instance learning with instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 958–977, May 2011.
[17] E. T. Frank and X. Xu, "Applying propositional learning algorithms to multi-instance data," University of Waikato, Hamilton, NZ, USA, 06/03 2003.
[18] J. Foulds and E. Frank, "Revisiting multiple-instance learning via embedded instance selection," in *Proc. Australasian Joint Conf. Artif. Intell.*, 2008, pp. 300–310.
[19] L. Dong, "A comparison of multi-instance learning algorithms," University of Waikato, Hamilton, NZ, USA, 2006.
[20] J. Wu, Z. Hong, S. Pan, X. Zhu, C. Zhang, and Z. Cai, "Multi-graph learning with positive and unlabeled bags," in *Proc. 7th SIAM Int. Conf. Data Mining*, 2014, pp. 217–225.
[21] W. J. Li and D. Y. yeung, "MILD: Multiple-instance learning via disambiguation," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 1, pp. 76–89, Jan. 2010.
[22] J. Wang, "Solving the multiple-instance problem: A lazy learning approach," in *Proc. 29th Int. Conf. Mach. Learn.*, 2000, pp. 1119–1125.
[23] Z. Geler, V. Kurbalija, M. Radovanović, and M. Ivanović, "Comparison of different weighting schemes for the kNN classifier on time-series data," *Knowl. Inf. Syst.*, vol. 48, no. 2, pp. 331–378, 2016.
[24] L. Bjerring and E. Frank, "Beyond trees: Adopting MITI to learn rules and ensemble classifiers for multi-instance data," in *Proc. Australasian Joint Conf. Artif. Intell.*, 2011, pp. 41–50.
[25] D. Nguyen, C. Nguyen, R. Hargraves, L. Kurgan, and K. Cios, "mi-DS: Multiple-instance learning algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 143–154, Feb. 2013.
[26] X. Qi and Y. Han, "Incorporating multiple SVMs for automatic image annotation," *Pattern Recogn.*, vol. 40, pp. 728–741, 2007.
[27] S. Ray and M. Craven, "Supervised versus multiple instance learning: An empirical comparison," in *Proc. 29th Int. Conf. Mach. Learn.*, 2005, pp. 697–704.
[28] X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," in *Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, 2004, pp. 272–281.
[29] V. Cheplygina, D. M. Tax, and M. Loog, "Multiple instance learning with bag dissimilarities," *Pattern Recogn.*, vol. 48, no. 1, pp. 264–275, 2015.
[30] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 1998, pp. 570–576.
[31] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2001, pp. 1073–1080.
[32] P. Auer and R. Ortner, "A boosting approach to multiple instance learning," in *Proc. Eur. Conf. Mach. Learn.*, 2004, pp. 63–74.
[33] D. Zhang, J. He, and R. Lawrence, "MI2LS: Multi-instance learning from multiple informationsources," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 149–157.
[34] Y. Xiao, B. Liu, Z. Hao, and L. Cao, "A similarity-based classification framework for multiple-instance learning," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 500–515, Apr. 2014.
[35] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.
[36] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 490–503.
[37] J. R. Foulds, "learning instance weight in multi-instance learning," Ph.D. dissertation, Department of Computer Science, University of Waikato, Hamilton, NZ, USA, 2008.
[38] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 416–431, Mar. 2006.
[39] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.
[40] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, Dec. 2004.
[41] S. Scott, J. Zhang, and J. Brown, "On generalized multiple-instance learning," *Int. J. Comput. Intell. Appl.*, vol. 05, no. 01, pp. 21–35, 2005.
[42] S. Boughorbel, J. P. Tarel, and N. Boujemaa, "The intermediate matching kernel for image local features," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2005, vol. 2, pp. 889–894.
[43] M.-L. Zhang and Z.-H. Zhou, "Multi-instance clustering with applications to multi-instance prediction," *Appl. Intell.*, vol. 31, no. 1, pp. 47–68, 2009.
[44] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proc. 29th Int. Conf. Mach. Learn.*, 2007, pp. 105–112.

[45] T. Gartner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proc. 29th Int. Conf. Mach. Learn.*, 2002, pp. 179–186.
[46] Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proc. 29th Int. Conf. Mach. Learn.*, 2007, pp. 1167–1174.
[47] D. M. Tax and R. P. Duin, "Learning curves for the analysis of multiple instance classifiers," in *Proc. Joint IAPR Int. Workshops Structural Syntactic Pattern Recog.*, 2008, pp. 724–733.
[48] J. Wu, X. Zhu, and C. Zhang, "Multi-instnace multi-graph dual embedding learning," in *Proc. IEEE Int. Conf. Data Mining*, 2013, pp. 827–836.
[49] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 561–568.
[50] X.-S. Wei, J. Wu, and Z.-H. Zhou, "Scalable multi-instance learning," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 1037–1042.
[51] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.
[52] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," in *Proc. Int. Conf. Adv. Visual Inf. Syst.*, 1999, pp. 509–516.
[53] Y. Li, S. Wang, Q. Tian, and X. Ding, "A boosting approach to exploit instance correlations for multi-instance classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2740–2747, Dec. 2016.
[54] A. Srinivasan, and S. H. Muggleton, "Mutagenesis: ILP experiments in a non-determinate biological domain," in *Proc. 4th Int. Workshop Inductive Logic Program.*, 1994, pp. 217–232.
[55] D. Michie, S. Muggleton, D. Page, and A. Srinivasan, "A new east-west challenge," Oxford University Computing Laboratory, Oxford, U.K., 1994.
[56] P. Reutemann, B. Pfahringer, and E. Frank, "A toolbox for learning from relational data with propositional and multi-instance learners," in *Proc. Australasian Joint Conf. Artif. Intell.*, 2004, pp. 1017–1023.
[57] J. He, H. Gu, and Z. Wang, "Bayesian multi-instance multi-label learning using gaussian process prior," *Mach. Learn.*, vol. 88, pp. 273–295, 2012.
[58] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews," in *Proc. 23rd Int. Conf. World Wide Web*, 2013, pp. 897–908.
[59] M. Kandemir and F. A. Hamprecht, "Instance label prediction by dirichlet process multiple instance learning," in *Proc. 26th Conf. Uncertainty Artif. Intell.*, 2014, pp. 380–389.
[60] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1289–1296.
[61] M. Haddoud, A. Mokhtari, T. Lecroq, and S. Abdeddaïm, "Combining supervised term-weighting metrics for SVM text classification with extended term representation," *Knowl. Inf. Syst.*, vol. 49, no. 3, pp. 909–931, 2016.
[62] X. S. Wei, J. Wu, and Z. H. Zhou, "Scalable algorithms for multi-instance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 975–987, Apr. 2017.

Jia Wu received the PhD degree in computer science from the University of Technology Sydney, Australia. He is currently a lecturer in the Department of Computing, Faculty of Science and Engineering, Macquarie University, Sydney. Prior to that, he was with the Centre for Artificial Intelligence, University of Technology Sydney. His research focuses on data mining and machine learning. Since 2009, he has published more than 20 refereed journal and conference papers (such as the *IEEE Transactions on Knowledge and Data Engineering*, and the *IEEE Transactions on Cybernetics*, *Pattern Recognition*, IJCAI, ICDM, SDM, and CIKM) in these areas.

Shirui Pan received the PhD degree in computer science from the University of Technology, Sydney (UTS). He is a research associate in the Centre for Artificial Intelligence (CAI), University of Technology, Sydney. His research interests include data mining and machine learning. To date, he has published more than 40 research papers in top-tier journals and conferences, including the *IEEE Transactions on Knowledge and Data Engineering*, the *IEEE Transactions on Cybernetics*, *Pattern Recognition*, IJCAI, ICDE, ICDM, SDM, and CIKM.

Xingquan Zhu (SM'12) received the PhD degree in computer science from Fudan University, China. He is an associate professor in the Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University. His research interests mainly include data mining, machine learning, bioinformatics, and computational advertising. Since 2010, he has published over 220 refereed journal and conference papers in these areas, and has won two Best Paper Awards and one Best Student Paper Award. He is an associate editor of the *IEEE Transactions on Knowledge and Data Engineering* (2008-2012, 2014-date), and an associate editor of the *ACM Transactions on Knowledge Discovery from Data* (2017-date). His a senior member of the IEEE.

Chengqi Zhang (SM'95) received the PhD degree from the University of Queensland, Brisbane, Australia, and the DSc degree (higher doctorate) from Deakin University, Geelong, Australia, in 1991 and 2002, respectively. Since December 2001, he has been a professor of information technology with the University of Technology, Sydney (UTS), Sydney, Australia, and he has been the director of the University of Technology, Sydney Priority Investment Research Centre for Quantum Computation and Intelligent Systems since April 2008. His research interests mainly focus on data mining and its applications. He is general co-chair of the Knowledge Discovery in Databases 2015 in Sydney, the local arrangements chair of IJCAI-2017 in Melbourne, a fellow of the Australian Computer Society, and a senior member of the IEEE.

Xindong Wu (F'11) received the PhD degree in artificial intelligence from the University of Edinburgh, Britain. He is a Yangtze River scholar in the School of Computer Science and Information Engineering, Hefei University of Technology, China, and a professor of computer science with the University of Louisiana at Lafayette. He is the steering committee chair of the IEEE International Conference on Data Mining (ICDM) and editor-in-chief of the *Knowledge and Information Systems (KAIS)*. He was the editor-in-chief of the *IEEE Transactions on Knowledge and Data Engineering (TKDE)* from 2005 to 2008. His research interests include data mining and big data analytics. He is a fellow of the IEEE and the AAAS.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.