Incorporating Diversity and Informativeness in Multiple-Instance Active Learning

Ran Wang, *Member, IEEE*, Xi-Zhao Wang, *Fellow, IEEE*, Sam Kwong, *Fellow, IEEE*, and Chen Xu

***Abstract*—Multiple-instance active learning (MIAL) is a paradigm to collect sufficient training bags for a multiple-instance learning (MIL) problem, by selecting and querying the most valuable unlabeled bags iteratively. Existing works on MIAL evaluate an unlabeled bag by its informativeness with regard to the current classifier, but neglect the internal distribution of its instances, which can reflect the diversity of the bag. In this paper, two diversity criteria, i.e., clustering-based diversity and fuzzy rough set based diversity, are proposed for MIAL by utilizing a support vector machine (SVM) based MIL classifier. In the first criterion, a kernel k-means clustering algorithm is used to explore the hidden structure of the instances in the feature space of the SVM, and the diversity degree of an unlabeled bag is measured by the number of unique clusters covered by the bag. In the second criterion, the lower approximations in fuzzy rough sets are used to define a new concept named dissimilarity degree, which depicts the uniqueness of an instance so as to measure the diversity degree of a bag. By incorporating the proposed diversity criteria with existing informativeness measurements, new MIAL algorithms are developed, which can select bags with both high informativeness and diversity. Experimental comparisons demonstrate the feasibility and effectiveness of the proposed methods.**

***Index Terms*—Clustering, diversity, fuzzy rough set, multiple-instance active learning (MIAL).**

I. Introduction

MULTIPLE-INSTANCE learning (MIL) [1] is a supervised learning problem that aims to construct a classification model on structured data. Different from traditional single-instance learning (SIL) that consists of individual samples, MIL has samples grouped into bags. A bag is decided as positive if at least one of its instances is positive, and is negative only if all of its instances are negative. The numbers of instances in different bags may differ a lot, and the instance-level labels are usually unknown. The objective of MIL is to learn a classifier that can accurately predict labels of new bags. Many real-world scenarios can be categorized as MIL problems. For instance, in image recognition, we consider an image as containing an object if any area of the image contains the object, and in text classification, we consider a text file as positive if any paragraph of the text file is positive.

Many solutions have been proposed for solving MIL problems, such as axis-parallel rectangles [2], k-nearest neighbor (kNN) based algorithms (i.e., Bayesian-kNN and Citation-kNN) [3], problem transformation method (i.e., MILES) [4], random walk process [5], graph kernel based method (i.e., MI-Graph) [6], genetic programming algorithm (i.e., G3P-MI) [7], combinatorial margin maximization formulation [8], random set framework [9], similarity-based framework [10], query-adaptive approach [11], and others [12], [13]. In this paper, we only focus on the learning methods based on the support vector machine (SVM) [14], whose idea is to generate an optimal separating hyperplane that can maximize the margin between two classes. Two SVM-based MIL algorithms have been proposed in [15] (i.e., mi-SVM and MI-SVM), which construct an SVM classifier by maximizing the margin between instances or between bags, and get the decision of a new bag by predicting the labels of its instances. Due to a simple implementation procedure and a high generalization capability, SVM-based MIL algorithms have been utilized in many application domains, such as music information retrieval [16], image annotation [17], [18], and human detection [19].

On the other hand, in many real-world problems, label acquisition is expensive due to the involvement of human efforts. Thus, active learning (AL) [20] becomes a commonly used scheme for collecting a sufficiently large labeled set by querying the informative unlabeled samples in an iterative manner. AL has been achieved for traditional single-instance settings with many successful sample selection criteria (e.g., inconsistency [21], fuzzy rough set [22], ambiguity [23], etc.), and it has been applied to many application domains (e.g., multilabel image classification [24], video instance retrieval [25], target ranking [25], imbalanced data classification [26], [27], etc.).

Manuscript received February 26, 2016; revised August 8, 2016 and February 3, 2017; accepted June 3, 2017. Date of publication June 20, 2017; date of current version November 29, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61402460, Grant 71371063, Grant 61472257, and Grant 61672443, in part by the Natural Science Foundation of SZU under Grant 2017060, in part by the Basic Research Project of Knowledge Innovation Program in Shenzhen under Grant JCYJ20150324140036825, in part by the Guangdong Provincial Science and Technology Plan Project under Grant 2013B040403005, in part by the HD Video R&D Platform for Intelligent Analysis and Processing in Guangdong Engineering Technology Research Centre of Colleges and Universities under Grant GCZX-A1409, and in part by the Hong Kong RGC General Research Fund under Grant 9042322 (CityU 11200116). *(Corresponding author: Xi-Zhao Wang.)*

R. Wang and C. Xu are with the College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, China (e-mail: wangran@szu.edu.cn; xuchen@szu.edu.cn).

X.-Z. Wang is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: xizhaowang@ieee.org).

S. Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: CSSAMK@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2017.2717803

However, in multiple-instance environments, AL is much less studied. Multiple-instance AL (MIAL) aims to make as few queries as possible for training bags, and constructs an MIL classifier to achieve high prediction accuracy on the labels of testing bags.

By taking mi-SVM or MI-SVM as the base classifier, margin-based MIAL strategy was first proposed in [28], which evaluates the informativeness of an unlabeled bag by aggregating the margin information of its instances. Later, the softmax model was proposed in [29], which transforms the output of an SVM into probabilistic form, and evaluates the informativeness of an unlabeled bag by its label uncertainty. Furthermore, the combinU [30] model and noisy-or model [31] were also used in a similar framework. Besides, Fisher information was used to evaluate the amount of information carried by a bag [32], and a multicriteria decision making system was proposed to select bags by making use of multiple conflicting criteria [33].

It is obvious that all the above-mentioned strategies try to select and query the most informative bags with regard to the current classifier. This is consistent with the central idea of traditional AL, however, is insufficient for MIAL, since it neglects the internal distribution and the relationship among different instances in a bag. For example, assume that two bags have obtained very similar uncertainty according to the current classifier, but the instances are located sparsely in one bag and densely in the other. Traditional methods will treat them as equally informative and select one randomly. However, intuitively, the bag with sparse distribution should be queried, since it can span the feature space and force the hyperplane to converge to the optimal one faster. This kind of information can be reflected by the diversity of a bag, which relies on the characteristics of the instance distribution, while it is independent of the current classifier. To the best of our knowledge, incorporating diversity to MIAL has not been investigated yet. Motivated by this fact, in this paper, we will propose two diversity criteria, and apply them to develop new MIAL schemes. The innovations and contributions of this paper are listed as follows.

1) We propose a clustering-based diversity (CBD) criterion for MIAL. A kernel k-means clustering algorithm is conducted on the instances of the candidate bags to explore the hidden structure of the instances in feature space. Then, a diversity index is calculated for each bag as the number of unique clusters it covers. The adoption of a fixed kernel guarantees that the clustering process is conducted in the same feature space of an SVM.
2) We propose a fuzzy rough set based diversity (FBD) criterion for MIAL. A new concept named dissimilarity degree is proposed based on the lower approximations in fuzzy rough sets, which measures the uniqueness of an instance in a bag. Then, the diversity degree of a bag is calculated by aggregating the dissimilarity degrees of its instances. Similarly, the adoption of kernel-based similarity relation makes it intrinsically compatible with the SVM.
3) We develop new MIAL strategies by incorporating the proposed diversity criteria. First, the informativeness of unlabeled bags are evaluated by the output decisions of the current classifier. Then, the most informative bags are retained as the selective candidates. Afterward, the diversity values of the candidate bags are computed according to the proposed criteria. Finally, the unlabeled bag with the highest diversity value is selected for manual labeling, it is then used to update the MIL classifier.
4) We generate new MIL datasets from the MNIST handwritten digits image recognition problem, and conduct extensive experimental comparisons on both existing MIL datasets and the generated datasets to validate the performance of the proposed MIAL strategies.

The remainder of this paper is organized as follows. In Section II, we introduce some background knowledge. In Section III, we present our motivation, then propose two diversity criteria and apply them to SVM-based MIAL. In Section IV, we conduct extensive experimental comparisons to show the feasibility and effectiveness of the proposed methods. Finally, conclusions are given in Section V.

II. Background Knowledge

In this section, we will introduce SVM-based MIL algorithms and present some preliminaries on SVM-based MIAL.

A. SVM-Based MIL

The traditional SVM on single-instance training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathcal{R}^d \times \{+1,-1\}$ is defined as $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$, where $\mathbf{w}$ is a d-dimensional vector and b is a constant. With the Lagrange method, the solution $(\mathbf{w}, b)$ can be derived by solving the following optimization problem:

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i,$$
$$\text{s.t.} \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n \tag{1}$$

where C is a tradeoff constant and ξ_i is the slack variable introduced to $\mathbf{x}_i$ for a soft-margin SVM. By incorporating kernel trick, the decision function is further represented as $f(\mathbf{x}) = \sum_{i=1}^n y_i\alpha_i\mathcal{K}(\mathbf{x}, \mathbf{x}_i) + b$, where α_i is the Lagrange multiplier of $\mathbf{x}_i$ and $\mathcal{K}(\cdot,\cdot)$ is a kernel function.

Given a multiple-instance training set $\mathbb{S} = \{(\mathcal{B}_i, y_i)\}_{i=1}^n$, we denote $\mathcal{B}_i = \{\mathcal{B}_{ij}\}_{j=1}^{n_i}$ as the ith bag with n_i instances, and $y_i \in \{+1,-1\}$ as the label of $\mathcal{B}_i$, where the instance-level labels y_{ij} are unknown. The goal is to construct a classification model that can accurately predict the labels of new bags. An illustration of multiple-instance dataset is given in Fig. 1.

The SVM has been extended to MIL by maximizing the margin between instances or between bags [15]. More specifically, mi-SVM is a maximum instance margin formulation inspired by the idea that all the instances of the negative bags are located in the negative half-space, and at least one instance of each positive bag is located in the positive half-space. The mi-SVM is

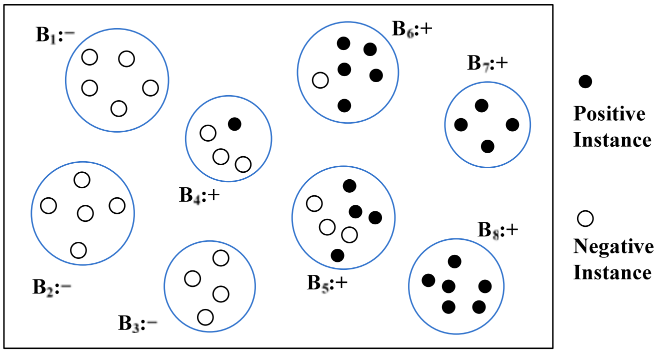


Fig. 1. Instances and bags in MIL.

formulated as

$$\begin{aligned}
&\min_{\{y_{ij}\}} \quad \min_{\mathbf{x},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i\sum_j \xi_{ij}, \\
&\text{s.t.} \quad \forall i,j: y_{ij}(\mathbf{w}^{\mathrm{T}}\mathcal{B}_{ij}+b) \geq 1-\xi_{ij}, \\
&\qquad \xi_{ij} \geq 0,\ y_{ij} \in \{+1,-1\}, \\
&\qquad \forall i \ \text{ s.t. } y_i = +1: \ \sum\nolimits_{\mathcal{B}_{ij}\in\mathcal{B}_i} \frac{y_{ij}+1}{2} \geq 1, \\
&\qquad \forall i \ \text{ s.t. } y_i = -1: \ y_{ij} = -1
\end{aligned} \tag{2}$$

where ξ_{ij} is the slack variable introduced to $\mathcal{B}_{ij}$.

On the other hand, MI-SVM is a maximum bag margin formulation inspired by the idea that each positive bag can be replaced by the most representative instance, which is defined as the instance that has the maximum decision value in a positive bag. The MI-SVM is formulated as

$$\begin{aligned}
&\min_{\mathbf{x},b,\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i, \\
&\text{s.t.} \quad \forall i: \ y_i \max_{j\in\mathcal{B}_i}(\mathbf{w}^{\mathrm{T}}\mathcal{B}_{ij}+b) \geq 1-\xi_i, \\
&\qquad \xi_i \geq 0
\end{aligned} \tag{3}$$

where ξ_i is the slack variable introduced to $\mathcal{B}_i$.

The heuristic algorithms of mi-SVM and MI-SVM are given in Appendix A.

B. SVM-Based MIAL

In this paper, we only deal with pool-based AL. In the MIL environment, pool-based AL starts by training a classifier with a small number of labeled bags. Afterward, it evaluates all the unlabeled bags in the selective pool, queries the most valuable bag, adds the bag to the training set, and updates the current classifier. This process repeats until a predefined stopping criterion is satisfied. Obviously, the key issue is to design an effective evaluation criterion for bags. The basic framework for SVM-based MIAL is described in Algorithm 1.

Algorithm 1: Basic Framework for SVM-based MIAL

```
Input:
    Labeled set $\mathbb{L} = \{(\mathcal{B}_i, y_i)\}_{i=1}^{l}$;
    Unlabeled pool $\mathbb{U} = \{\mathcal{B}_i\}_{i=l+1}^{l+u}$;
    Parameters for training SVM.
Output:
    SVM solution $(\mathbf{w}, b)$.
Train mi-SVM or MI-SVM on $\mathbb{L}$ to get SVM solution $(\mathbf{w}, b)$;
while $\mathbb{U}$ is not empty do
    if stop criterion is met then
        return $(\mathbf{w}, b)$;
    else
        Calculate the informativeness of each $\mathcal{B}_i \in \mathbb{U}$, denoted as $\mathcal{I}(\mathcal{B}_i)$;
        Select $\mathcal{B}^* = \text{argmax}_{\mathcal{B}_i\in\mathbb{U}}\mathcal{I}(\mathcal{B}_i)$;
        Query the label of $\mathcal{B}^*$, denoted by $y^*$;
        Let $\mathbb{U} = \mathbb{U} \setminus \mathcal{B}^*$, and $\mathbb{L} = \mathbb{L} \cup (\mathcal{B}^*, y^*)$;
        Update SVM solution $(\mathbf{w}, b)$ based on new $\mathbb{L}$;
    end
end
return $(\mathbf{w}, b)$.
```

C. Informativeness Measurements for Unlabeled Bags

Existing solution to assessing an unlabeled bag for SVM-based MIL algorithms is to aggregate the informativeness of its instances. Assuming that $\mathcal{B}_{ij}$ is the jth instance in the ith unlabeled bag. Given an SVM classifier $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b$, the decision value of $\mathcal{B}_{ij}$ is calculated as $f(\mathcal{B}_{ij}) = \mathbf{w}^{\mathrm{T}}\phi(\mathcal{B}_{ij}) + b$, and its conditional probabilities can be evaluated by the logistic function [34]:

$$\begin{cases} \mathcal{P}(y_{ij} = +1|\mathcal{B}_{ij}) = \frac{1}{1+\exp(-f(\mathcal{B}_{ij}))} \\ \mathcal{P}(y_{ij} = -1|\mathcal{B}_{ij}) = 1 - \mathcal{P}(y_{ij} = +1|\mathcal{B}_{ij}). \end{cases} \tag{4}$$

Furthermore, the uncertainty of $\mathcal{B}_{ij}$ can be computed as

$$u(\mathcal{B}_{ij}) = -\sum\nolimits_{y_{ij}=\pm 1} \mathcal{P}(y_{ij}|\mathcal{B}_{ij}) log\mathcal{P}(y_{ij}|\mathcal{B}_{ij}). \tag{5}$$

1) *Bag margin:* According to Liu *et al.* [28], the most straightforward way to evaluate the informativeness of $\mathcal{B}_i$ is the minimum instance margin $\mathcal{I}_m(\mathcal{B}_i)$ or the average instance margin $\mathcal{I}_a(\mathcal{B}_i)$:

$$\begin{cases} \mathcal{I}_m(\mathcal{B}_i) = 1/\min_{\mathcal{B}_{ij}\in\mathcal{B}_i}|f(\mathcal{B}_{ij})| \\ \mathcal{I}_a(\mathcal{B}_i) \ = \sqrt{n_i/\sum_{\mathcal{B}_{ij}\in\mathcal{B}_i}|f(\mathcal{B}_{ij})|^2}. \end{cases} \tag{6}$$

2) *Softmax model:* A softmax model is proposed in [29] to approximate the conditional probabilities of unlabeled bags. For a given set of values $x_1, \ldots, x_n$, the softmax approximation is given as

$$\text{softmax}_\alpha(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \cdot e^{\alpha\cdot x_i} \Big/ \sum_{i=1}^{n} e^{\alpha\cdot x_i} \tag{7}$$

where α is a parameterized constant. Then, the conditional probabilities of $\mathcal{B}_i$ are evaluated as

$$\begin{cases} \mathcal{P}(y_i = +1|\mathcal{B}_i) = & \operatorname{softmax}_\alpha(\mathcal{P}(y_{i1} = +1|\mathcal{B}_{i1}), \\ & \quad \ldots, \mathcal{P}(y_{in_i} = +1|\mathcal{B}_{in_i})) \\ \mathcal{P}(y_i = -1|\mathcal{B}_i) = & 1 - \mathcal{P}(y_i = +1|\mathcal{B}_i) \end{cases} \tag{8}$$

where $\operatorname{softmax}_\alpha$ is given in (7) to approximate the positive probability of the bag $\mathcal{P}(y_i = +1|\mathcal{B}_i)$ by the conditional probabilities of the instances $\mathcal{P}(y_{i1} = +1|\mathcal{B}_{ij}), j = 1, \ldots, n_i$. Finally, the informativeness of $\mathcal{B}_i$ is computed as

$$\mathcal{I}(\mathcal{B}_i) = -\sum_{y_i=\pm 1} \mathcal{P}(y_i|\mathcal{B}_i)\log\mathcal{P}(y_i|\mathcal{B}_i). \tag{9}$$

3) *CombinU model:* The combinU model [30] is an alternative to the softmax model by making the softmax approximation of the instance uncertainties, which evaluates the informativeness of $\mathcal{B}_i$ as

$$\mathcal{I}(\mathcal{B}_i) = \operatorname{softmax}_\alpha(u(\mathcal{B}_{i1}), \ldots, u(\mathcal{B}_{in_i})) \tag{10}$$

where $\operatorname{softmax}_\alpha$ is given in (7) to approximate the uncertainty of the bag $\mathcal{I}(\mathcal{B}_i)$ by the uncertainties of the instances $u(\mathcal{B}_{ij}), j = 1, \ldots, n_i$.

4) *Noisy-or model:* The noisy-or model [31] is a nonparametric structure that can capture the nondeterministic interaction between different causes of an effect. By utilizing this model, the conditional probabilities of $\mathcal{B}_i$ can be evaluated as

$$\begin{cases} \mathcal{P}(y_i = +1|\mathcal{B}_i) = 1 - \prod_{\mathcal{B}_{ij}\in\mathcal{B}_i}(1 - \mathcal{P}(y_{ij} = +1|\mathcal{B}_{ij})) \\ \mathcal{P}(y_i = -1|\mathcal{B}_{ij}) = \prod_{\mathcal{B}_{ij}\in\mathcal{B}_i}(1 - \mathcal{P}(y_{ij} = +1|\mathcal{B}_{ij})) \end{cases} \tag{11}$$

and the informativeness of $\mathcal{B}_i$ is computed by (9).

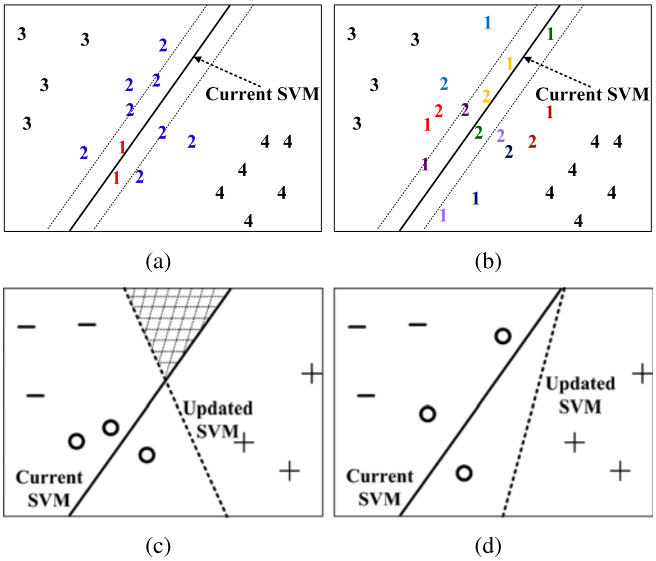
5) *Fisher information:* Fisher information [35] has been successfully used to measure the amount of information that a batch of samples carries with regard to a classification model [36]. The Fisher information matrix of n i.i.d samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with distribution $q(\mathbf{x})$ and classification model $p(y|\mathbf{x}, \theta)$ is defined as

$$I_{q(\mathbf{x})}(\theta) = -\int q(\mathbf{x})d\mathbf{x}\int p(y|\mathbf{x}, \theta)\frac{\partial^2}{\partial\theta^2}\log p(y|\mathbf{x}, \theta)d\mathbf{x} \tag{12}$$

where θ denotes the model parameters. This model has been applied to MIAL [32], [33], [37] for measuring the informativeness of unlabeled bag $\mathcal{B}_i$ by an effective approximation:

$$\begin{aligned} &\mathcal{I}(\mathcal{B}_i) \\ &= -tr\Big(\sum_{y_i=\pm 1} p(y_i|\mathcal{B}_i, \theta)\frac{\partial^2}{\partial\theta^2}\log p(y_i|\mathcal{B}_i, \theta)\Big) \\ &= (\mathcal{P}_i^-/\mathcal{P}_i^+) \times \sum_{j\in\mathcal{B}_i}(\mathcal{P}_{ij}^+ \times \phi(\mathcal{B}_{ij}))^{\mathrm{T}}\sum_{j\in\mathcal{B}_i}(\mathcal{P}_{ij}^+ \times \phi(\mathcal{B}_{ij})) \\ &= (\mathcal{P}_i^-/\mathcal{P}_i^+) \times \sum_{j\in\mathcal{B}_i}\sum_{q\in\mathcal{B}_i}\mathcal{P}_{ij}^+ \times \mathcal{P}_{iq}^+ \times \mathcal{K}(\mathcal{B}_{ij}, \mathcal{B}_{iq}) \end{aligned} \tag{13}$$

where ϕ denotes a kernel mapping, $\mathcal{K}(\cdot,\cdot)$ is the kernel function, $\mathcal{P}_i^+$, $\mathcal{P}_i^-$, and $\mathcal{P}_{ij}^+$ denote $\mathcal{P}(y_i = +1|\mathcal{B}_i)$, $\mathcal{P}(y_i = -1|\mathcal{B}_i)$, and $\mathcal{P}(y_{ij} = +1|\mathcal{B}_{ij})$, respectively.



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

Fig. 2. Investigation on unlabeled bags for an SVM classifier.

III. Incorporating Diversity in MIAL

In this section, we will present our motivation in detail, then develop two diversity-based MIAL algorithms.

A. Motivation

The evaluation of bags in MIAL is more difficult than that of samples in traditional AL due to the complex combinations of instances in bags. In general, there are two basic characteristics of the bags in MIAL: 1) the number of instances in different bags may differ a lot; and 2) the instances in a bag may have various distributions.

It is well known that in traditional SVM-based AL, the samples closer to the current decision boundary are more informative. In fact, all the measurements of bags introduced in Section II-C are based on this criterion. Having this premise, in Fig. 2, we further analyze the influence of the above-mentioned two characteristics on measuring the informativeness of unlabeled bags for an SVM classifier, which have not be well addressed by existing measurements.

1) In Fig. 2(a), it is obvious that Bag 1 and Bag 2 are more informative than Bag 3 and Bag 4. Existing methods prefer Bag 1 than Bag 2, since the instances in Bag 1 are closer to the SVM hyperplane. However, Bag 2 might be more valuable, since it contains much more information, although the instances are less informative than those in Bag 1.
2) In Fig. 2(b), both Bag 1 and Bag 2 contain eight instances. We use the same color to represent the same level of informativeness. As a result, the informativeness of the two bags are similar, and existing methods will select one randomly. However, intuitively, Bag 1 is more valuable than *Bag* 2, since the distribution of instances in Bag 1 is sparser, which can span the feature space and induce a better SVM classifier. We give a further illustration in Fig. 2(c) and (d). Suppose that during a learning iteration, the selected bag is negative. In MIL, a bag is determined as negative only if all the instances are negative. If the bag has a dense distribution, as shown in Fig. 2(c), the updated SVM hyperplane will be biased by this small area. In this case, the negative instances located in other areas will be wrongly classified with high probability [such as the highlighted area in Fig. 2(c)], thus negative bags might be wrongly classified. However, if the bag has a sparse distribution, as shown in Fig. 2(d), this problem can be avoided to some extent.

In order to evaluate unlabeled bags more effectively, we propose a new criterion named diversity. Different from the criterion of informativeness, diversity is independent of the current classifier, and is just decided by the properties of the bag. In Sections III-B and III-C, we will develop two diversity measurements by applying the techniques of kernel-based clustering and fuzzy rough sets. It is noteworthy that these two techniques have a common feature, which make them suitable for solving the SVM problem. More specifically, a kernel function measures the similarity between two samples in the feature space of an SVM, it can also serve as the distance measurement in kernel-based clustering and the fuzzy similarity relation in fuzzy rough sets. The same kernel function guarantees that all the learning processes are conducted in the same feature space. That is to say, the kernel technique makes them intrinsically compatible with the SVM.

B. MIAL with CBD

The kernel k-means algorithm performs the clustering in a higher dimensional feature space instead of the original space. It includes several key steps:

1) randomly initialize k clusters;

2) compute the distance of each sample to the center of each cluster in kernel space, and assign it to the closest cluster; and

3) repeat step 2 until the clusters have no change.

Given an unlabeled sample set $\{\mathbf{x}_i\}_{i=1}^N \subset \mathcal{R}^d$, we denote $\phi(\mathbf{x}_i)$ as the data point of $\mathbf{x}_i$ in kernel space, $\mathcal{C}_\nu$ as the νth cluster, where $\nu = 1, \ldots, k$, μ_ν as the center of cluster $\mathcal{C}_\nu$, and $\mathcal{C}^{(i)}$ as the cluster index of $\mathbf{x}_i$. Obviously, the center of cluster $\mathcal{C}_\nu$ in kernel space can be computed as

$$\mu_\nu = \frac{\sum_{\mathbf{x}_i \in \mathcal{C}_\nu} \phi(\mathbf{x}_i)}{|\mathcal{C}_\nu|}. \tag{14}$$

Algorithm 2: Kernel k-mean Algorithm

```
Input:
    Unlabeled data set $\{\mathbf{x}_i\}_{i=1}^N \subset \mathcal{R}^d$; Number of clusters $k$.
Output:
    Cluster indices $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(N)}$.
Randomly assign the cluster indices $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(N)}$ from $\{1, \ldots, k\}$;
repeat
    for each $\mathbf{x}_i$ do
        Compute $D^2(\phi(\mathbf{x}_i), \mu_\nu)$ by Eq. (15) where $\nu = 1, \ldots, k$;
        Let $\mathcal{C}^{(i)} = \text{argmin}_{\nu=1,\ldots,k} D^2(\phi(\mathbf{x}_i), \mu_\nu)$;
    end
until $\mathcal{C}^{(i)}, \ldots, \mathcal{C}^{(N)}$ have no change;
return $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(N)}$.
```

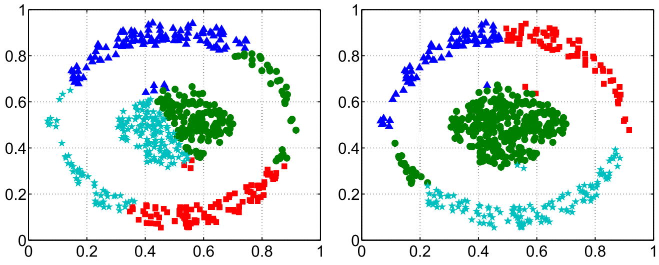
Since the concrete form of ϕ is unknown with regard to many kernels, it is hard to get the explicit expression of either $\phi(\mathbf{x}_i)$ or μ_ν. Similar to the SVM, the kernel trick can be used to express the inner product of feature space as a kernel function $\mathcal{K} : \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$. Thus, instead of getting the absolute location of a sample in kernel space, we can directly compute its distance to other samples. In this case, the distance between $\phi(\mathbf{x}_i)$ and cluster center μ_ν can be computed as

$$\begin{aligned} D^2(\phi(\mathbf{x}_i), \mu_\nu) &= ||\phi(\mathbf{x}_i) - \mu_\nu||^2 \\ &= ||\phi(\mathbf{x}_i) - \frac{1}{|\mathcal{C}_\nu|} \sum_{\mathbf{x}_j \in \mathcal{C}_\nu} \phi(\mathbf{x}_j)||^2 \\ &= \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{|\mathcal{C}_\nu|} \sum_{\mathbf{x}_j \in \mathcal{C}_\nu} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \\ &\quad + \frac{1}{|\mathcal{C}_\nu|^2} \sum_{\mathbf{x}_j \in \mathcal{C}_\nu} \sum_{\mathbf{x}_q \in \mathcal{C}_\nu} \mathcal{K}(\mathbf{x}_j, \mathbf{x}_q) \end{aligned} \tag{15}$$

where || || is defined as the Euclidian norm, i.e., the length of a vector.

As a result, kernel k-means clustering is described in Algorithm 2. It probes the hidden structure of the data in feature space, explores the relative location information of the instances, and groups together the similar instances from a spatial perspective. Fig. 3 demonstrates a set of clustering results by the k-means algorithm and kernel k-means algorithm. As shown in Fig. 3(b), the kernel k-means algorithm has a higher capability in handling nonlinearly separable case by transforming the instances into kernel space. According to this investigation, the kernel k-means algorithm might be more effective in handling problems with the SVM.

It is noteworthy that the kernel k-means algorithm needs the number of clusters k as an input. One solution is to set k as the number of instances in a bag, which guarantees that the instances in a bag will be grouped into different clusters if all of the instances have low similarities.



(a) (b)

Fig. 3. Illustrative examples for (a) k-means clustering and (b) kernel k-means clustering.

Algorithm 3: MIAL-CBD

```
Input:
    Labeled set $\mathbb{L} = \{(\mathcal{B}_i, y_i)\}_{i=1}^{l}$;
    Unlabeled pool $\mathbb{U} = \{\mathcal{B}_i\}_{i=l+1}^{l+u}$;
    Number of clusters $k$;
    Number of candidates $m$;
    Parameters for training SVM.
Output:
    SVM solution $(\mathbf{w}, b)$.
Train mi-SVM or MI-SVM on $\mathbb{L}$ to get SVM solution $(\mathbf{w}, b)$;
while $\mathbb{U}$ is not empty do
    if stop criterion is met then
        return $(\mathbf{w}, b)$;
    else
        Calculate $\mathcal{I}(\mathcal{B}_i)$ of each $\mathcal{B}_i \in \mathbb{U}$;
        Let $\mathbb{U}^* \in \mathbb{U}$ contain $m$ most informative bags;
        Let $\mathbb{X} = \{\mathcal{B}_{ij} | \mathcal{B}_{ij} \in \mathbb{U}^*\}$;
        Call Algorithm 2 on $\mathbb{X}$, denote $\mathcal{C}^{(ij)} \in \{1, \dots, k\}$ as
        the cluster index of $\mathcal{B}_{ij}$;
        for each $\mathcal{B}_i \in \mathbb{U}^*$ do
            Let $\mathcal{C}(\mathcal{B}_i)$ be the number of unique indices in
            $\{\mathcal{C}^{(ij)} | \mathcal{B}_{ij} \in \mathcal{B}_i\}$;
        end
        Select $\mathcal{B}^* = \mathrm{argmax}_{\mathcal{B}_i \in \mathbb{U}^*} \mathcal{C}(\mathcal{B}_i)$;
        Query the label of $\mathcal{B}^*$, denoted by $y^*$;
        Let $\mathbb{U} = \mathbb{U} \setminus \mathcal{B}^*$, and $\mathbb{L} = \mathbb{L} \cup (\mathcal{B}^*, y^*)$;
        Update SVM solution $(\mathbf{w}, b)$ based on new $\mathbb{L}$;
    end
end
return $(\mathbf{w}, b)$.
```

In SVM-based MIAL, we can first rank the unlabeled bags according to an informativeness criterion. Then, the top-ranked bags are retained as the selective candidates. Afterward, Algorithm 2 is conducted on the instances of the selective candidates, and the diversity of a candidate can be calculated as the number of unique clusters it covers. Finally, the MIAL algorithm with CBD is described in Algorithm 3.

C. MIAL with FBD

Fuzzy rough sets [38], as the generalizations of crisp rough sets, are popular tools for handling data with vagueness and uncertainty, with the ability of dealing with mixed types of features [39]–[41]. A fuzzy rough set is defined by two fuzzy sets, i.e., lower and upper approximations. In traditional classification problems, they can be used to describe the maximum and minimum membership degrees of a sample belonging to different decision classes. In this section, we explore their potentials for measuring the uniqueness of an instance in a bag in the MIL environment.

Given $I = [0, 1]$, we denote $\mathcal{T} : I^2 \to I$ and $\mathcal{S} : I^2 \to I$ as a dual pair of triangular norm (t-norm) and triangular conorm (t-conorm). Assume that U is a nonempty universe of discourse, D is a fuzzy subset of U, i.e., $D \subset \mathcal{F}(U)$, and R is a fuzzy similarity relation on the cardinal product of U that is reflexive, symmetric, and transitive. Then, the most general case of fuzzy rough set is based on a pair of t-norm and t-conorm:

$$\begin{cases} \overline{R_{\mathcal{T}}} D(\mathbf{x}) = \sup_{\mathbf{u} \in U} \mathcal{T}(R(\mathbf{x}, \mathbf{u}), D(\mathbf{u})) \\ \underline{R_{\mathcal{S}}} D(\mathbf{x}) = \inf_{\mathbf{u} \in U} \mathcal{S}(\mathcal{N}(R(\mathbf{x}, \mathbf{u})), D(\mathbf{u})) \end{cases} \tag{16}$$

where $\mathcal{N}$ is a negator, i.e., a decreasing mapping $I \to I$ that satisfies $\mathcal{N}(0) = 1$ and $\mathcal{N}(1) = 0$. In this paper, we always adopt the standard negator $\mathcal{N}_s(\alpha) = 1 - \alpha$.

A more commonly used fuzzy rough set is based on the residual implication θ and its dual σ [40], which are defined as follows:

$$\begin{cases} \theta(a, b) = \sup\{c \in [0, 1] : \mathcal{T}(a, c) \leq b\} \\ \sigma(a, b) = \inf\{c \in [0, 1] : \mathcal{S}(a, c) \geq b\}. \end{cases} \tag{17}$$

It is easy to prove that $\sigma(a, b) = 1 - \theta(1 - a, 1 - b)$ relative to the same t-norm. Accordingly, the fuzzy rough set based on θ and σ is defined as

$$\begin{cases} \overline{R_{\sigma}} D(\mathbf{x}) = \sup_{\mathbf{u} \in U} \sigma(\mathcal{N}(R(\mathbf{x}, \mathbf{u})), D(\mathbf{u})) \\ \underline{R_{\theta}} D(\mathbf{x}) = \inf_{\mathbf{u} \in U} \theta(R(\mathbf{x}, \mathbf{u}), D(\mathbf{u})). \end{cases} \tag{18}$$

In a previous work [22], we have proposed a concept named consistence degree based on (18) to depict the minimum requirement of sample $\mathbf{x}$ belonging to its decision class. Assume that $\mathrm{FD} = (U, C \cup D)$ is a fuzzy decision table, where U is a nonempty universe of discourse, C is a set of conditional attributes with at least one fuzzy attribute, and D is a decision attribute. Given two samples $\mathbf{x}, \mathbf{y} \in \mathrm{FD}$, we let $[\mathbf{x}]_D(\mathbf{y}) = 1$ if $\mathbf{x}$ and $\mathbf{y}$ have the same decision attribute and $[\mathbf{x}]_D(\mathbf{y}) = 0$, otherwise. Then, the consistence degree of sample $\mathbf{x}$ is defined as Definition 1.

Definition 1 ([22]): (Consistence degree) Given a sample $\mathbf{x}$ in the fuzzy decision table $\mathrm{FD} = \{U, C \cup D\}$, the consistence degree of $\mathbf{x}$ in FD is defined as

$$\mathrm{Con}_C(D)(\mathbf{x}) = \inf_{\mathbf{u} \in U} \theta(R(\mathbf{x}, \mathbf{u}), [\mathbf{x}]_D(\mathbf{u})). \tag{19}$$

Suppose $\mathbf{y}$ is a sample distinct from $\mathbf{x}$, it has been proved in [22] that

1) if $\theta(R(\mathbf{x}, \mathbf{y}), 0) < \mathrm{Con}_C(D)(\mathbf{x})$, then $\mathbf{x}$ and $\mathbf{y}$ always have the same label;
2) if $\theta(R(\mathbf{x}, \mathbf{y}), 0) \geq \mathrm{Con}_C(D)(\mathbf{x})$, then $\mathbf{x}$ and $\mathbf{y}$ may have different labels.

Given a value $\eta > \mathrm{Con}_C(D)(\mathbf{x})$. If $\theta(R(\mathbf{x}, \mathbf{y}), 0) < \eta$, it is possible that $\theta(R(\mathbf{x}, \mathbf{y}), 0) \geq \mathrm{Con}_C(D)(\mathbf{x})$, which cannot guarantee $\mathbf{x}$ and $\mathbf{y}$ having the same label. This statement holds for every sample in U. Thus, $\mathrm{Con}_C(D)(\mathbf{x})$ can be treated as the maximum value to guarantee $\mathbf{x}$ having the identical decision with another sample in U. Borrowing this idea, we propose a converse definition named dissimilarity degree, to measure the uniqueness of an instance in a bag in the MIL environment. Suppose that $\mathbb{U}$ is the unlabeled pool of an MIAL problem, $\mathbf{x}$ and $\mathbf{y}$ are two distinct instances in $\mathbb{U}$, let

$$[\mathbf{x}]_{\mathbb{U}}(\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ belong to the same bag} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Then, the dissimilarity degree of an instance $\mathbf{x}$ in a bag $\mathcal{B}$ is proposed as Definition 2.

Definition 2: (Dissimilarity degree) Given an instance $\mathbf{x}$ in the unlabeled pool $\mathbb{U}$ of an MIAL problem, the dissimilarity degree of $\mathbf{x}$ in any possible unlabeled bag $\mathcal{B}$ is defined as

$$\mathrm{Dis}_{\mathcal{B}}(\mathbf{x}) = \inf_{\mathbf{y}\in\{\mathcal{B}\backslash\mathbf{x}\}} \theta(R(\mathbf{x},\mathbf{y}), \mathcal{N}([\mathbf{x}]_{\mathbb{U}}(\mathbf{y}))). \tag{21}$$

From Definitions 1 and 2, we can find two distinctions between consistence degree and dissimilarity degree.

1) The consistence degree is defined for a labeled sample in a decision table in the SIL environment, which contains samples from different classes, whereas the dissimilarity degree is defined for an instance in a bag in the MIL environment, both the instance and the bag are unlabeled.
2) The consistence degree is an operational result of the similarity relation $R(\mathbf{x},\mathbf{y})$ and the consistency $[\mathbf{x}]_D(\mathbf{y})$ between two samples, which depicts the membership of a sample in its decision class, whereas the dissimilarity degree is an operational result of the similarity relation $R(\mathbf{x},\mathbf{y})$ and the inconsistency $\mathcal{N}([\mathbf{x}]_{\mathbb{U}}(\mathbf{y}))$ between two instances, which is a converse concept that depicts the uniqueness of an instance in a bag.

We further analyze the characteristics of the dissimilarity degree by applying the most commonly used similarity relation, i.e., Gaussian kernel based fuzzy similarity relation $R_G(\mathbf{x},\mathbf{y}) = \exp(-||\mathbf{x}-\mathbf{y}||^2/2\sigma^2)$. According to Hu *et al.* [42], $R_G(\mathbf{x},\mathbf{y})$ is reflexive, symmetric, and $\mathcal{T}_{\cos}$-transitive, where the pair of residuated implicators is defined as

$$\begin{cases} \mathcal{T}_{\cos}(a,b) = \max\{ab - \sqrt{1-a^2}\sqrt{1-b^2}, 0\} \\ \theta_{\cos}(a,b) = \begin{cases} 1 & \text{if } a \le b \\ ab + \sqrt{(1-a^2)(1-b^2)} & \text{if } a > b. \end{cases} \end{cases} \tag{22}$$

Accordingly, the dissimilarity degree is derived as

$$\mathrm{Dis}_{\mathcal{B}}(\mathbf{x}) = \inf_{\mathbf{y}\in\{\mathcal{B}\backslash\mathbf{x}\}} \theta_{\cos}(R_G(\mathbf{x},\mathbf{y}), \mathcal{N}([\mathbf{x}]_{\mathbb{U}}(\mathbf{y}))). \tag{23}$$

By applying (22) to (23), we have the following:

1) when $\mathbf{x} \notin \mathcal{B}$

$$\mathrm{Dis}_{\mathcal{B}}(\mathbf{x}) = \inf_{\mathbf{y}\in\mathcal{B}} \theta_{\cos}(R_G(\mathbf{x},\mathbf{y}), 1) = 1;$$

2) when $\mathbf{x} \in \mathcal{B}$

$$\begin{aligned} \mathrm{Dis}_{\mathcal{B}}(\mathbf{x}) &= \inf_{\mathbf{y}\in\{\mathcal{B}\backslash\mathbf{x}\}} \theta_{\cos}(R_G(\mathbf{x},\mathbf{y}), 0) \\ &= \inf_{\mathbf{y}\in\{\mathcal{B}\backslash\mathbf{x}\}} \sqrt{1-\exp^2(-||\mathbf{x}-\mathbf{y}||^2/2\sigma^2)}. \end{aligned}$$

From the above, it is concluded that the dissimilarity degree of $\mathbf{x}$ in $\mathcal{B}$ is always 1 if $\mathbf{x} \notin \mathcal{B}$. However, the dissimilarity degree of $\mathbf{x}$ in $\mathcal{B}$ is decided by its closest neighbor if $\mathbf{x} \in \mathcal{B}$, furthermore, the closer the closest neighbor is, the smaller the dissimilarity degree will be.

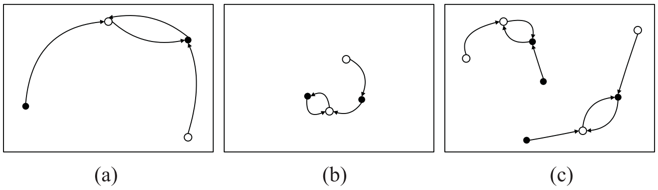


Fig. 4. Computing the dissimilarity degree of instances in a bag. (a) Bag 1, (b) Bag 2, and (c) Bag 3.

We further assume that $\mathbf{x} = \mathcal{B}_{ij}$ is the jth instance in the ith unlabeled bag, and $\mathbf{y}$ is an instance without bag information, then, we have the following theorem.

Theorem 1: Given two distinct instances $\mathcal{B}_{ij}$ and $\mathbf{y}$ in the unlabeled pool $\mathbb{U}$, if $\theta(R(\mathcal{B}_{ij},\mathbf{y}),0) < \mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})$, then $[\mathcal{B}_{ij}]_{\mathbb{U}}(\mathbf{y}) = 0$.

Proof: We prove it by contradiction. Assume $[\mathcal{B}_{ij}]_{\mathbb{U}}(\mathbf{y}) = 1$, then $\mathcal{N}([\mathcal{B}_{ij}]_{\mathbb{U}}(\mathbf{y})) = 0$. We have

$$\begin{aligned} \mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij}) &= \inf_{\mathbf{z}\in\{\mathcal{B}_i\backslash\mathcal{B}_{ij}\}} \theta(R(\mathcal{B}_{ij},\mathbf{z}), \mathcal{N}([\mathcal{B}_{ij}]_{\mathbb{U}}(\mathbf{z}))) \\ &\le \theta(R(\mathcal{B}_{ij},\mathbf{y}), \mathcal{N}([\mathcal{B}_{ij}]_{\mathbb{U}}(\mathbf{y}))) \\ &= \theta(R(\mathcal{B}_{ij},\mathbf{y}), 0). \end{aligned}$$

This contradicts the given condition of $\theta(R(\mathcal{B}_{ij},\mathbf{y}),0) < \mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})$, thus we get $[\mathcal{B}_{ij}]_{\mathbb{U}}(\mathbf{y}) = 0$. ■

According to Theorem 1, $\mathbf{y}$ is impossible to belong to $\mathcal{B}_i$ if $\theta(R(\mathcal{B}_{ij},\mathbf{y}),0) < \mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})$, and may belong to $\mathcal{B}_i$, otherwise. That is to say, $\mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})$ is the maximum value to guarantee the uniqueness of instance $\mathcal{B}_{ij}$ in bag $\mathcal{B}_i$. In other words, $\mathcal{B}_{ij}$ is highly different from other instances when the value of $\mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})$ is large, and may be similar to other instances when the value of $\mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})$ is small. Holding this argument, we propose a new concept named diversity degree as Definition 3 to measure the internal diversity among the instances in a given bag.

Definition 3: (Diversity degree) Given an unlabeled bag $\mathcal{B}_i = \{\mathcal{B}_{ij}\}_{j=1}^{n_i} \in \mathbb{U}$, the diversity of $\mathcal{B}_i$ is defined as

$$\mathrm{Div}(\mathcal{B}_i) = \sum_{\mathcal{B}_{ij}\in\mathcal{B}_i} \mathrm{Dis}_{\mathcal{B}_i}(\mathcal{B}_{ij})/|\mathcal{B}_i|. \tag{24}$$

Fig. 4 shows some illustrations on measuring the dissimilarity degree of instances and the diversity degree of bag. In this figure, each arrow represents that the dissimilarity degree of the start-point instance is decided by its distance to the end-point instance. According to (23), a larger distance will lead to a higher dissimilarity degree. Obviously, Bag 1 is the most diverse bag, since all the instances in this bag have higher dissimilarity degree. Finally, the MIAL algorithm with FBD is described in Algorithm 4.

D. Relationship Between CBD and FBD

So far, it is difficult to give a theoretical proof on the relationship between CBD and FBD. However, it is possible to give some intuitive explanations. Suppose that the same function is adopted as the kernel function in the CBD and the fuzzy similarity relation in the FBD. Then, the calculation of the CBD and

TABLE I
COMPARATIVE METHODS

Index	Method	Algorithm	Informativeness Measurement	Diversity Measurement
1	Random	1	Random sampling	None
2	SVMactive	1	Minimum instance margin [$\mathcal{I}_m$ in (6)]	None
3	BagMargin	1	Average instance margin [$\mathcal{I}_a$ in (6)]	None
4	SoftMax	1	Softmax model [see (8) and (9)]	None
5	CombinU	1	CombinU model [see (10)]	None
6	NoisyOr	1	Noisy-or model [see (11) and (9)]	None
7	Fisher	1	Fisher information [see (13)]	None
8	SVMactive + CBD	3	Minimum instance margin [$\mathcal{I}_m$ in (6)]	Clustering based
9	SVMactive + FBD	4	Minimum instance margin [$\mathcal{I}_m$ in (6)]	Fuzzy rough set based
10	NoisyOr + CBD	3	Noisy-or model [see (11) and (9)]	Clustering based
11	NoisyOr+FBD	4	Noisy-or model [see (11) and (9)]	Fuzzy rough set based
12	Fisher + CBD	3	Fisher information [see (13)]	Clustering based
13	Fisher + FBD	4	Fisher information [see (13)]	Fuzzy rough set based

```
Algorithm 4: MIAL-FBD
Input:
    Labeled set $\mathbb{L} = \{(\mathcal{B}_i, y_i)\}_{i=1}^{l}$;
    Unlabeled pool $\mathbb{U} = \{\mathcal{B}_i\}_{i=l+1}^{l+u}$;
    Number of candidates $m$;
    Parameters for training SVM.
Output:
    SVM solution $(\mathbf{w}, b)$.
Train mi-SVM or MI-SVM on $\mathbb{L}$ to get SVM solution $(\mathbf{w}, b)$;
while $\mathbb{U}$ is not empty do
    if stop criterion is met then
        return $(\mathbf{w}, b)$;
    else
        Calculate $\mathcal{I}(\mathcal{B}_i)$ of each $\mathcal{B}_i \in \mathbb{U}$;
        Let $\mathbb{U}^* \in \mathbb{U}$ contain $m$ most informative bags;
        for each $\mathcal{B}_i \in \mathbb{U}^*$ do
            Calculate its diversity degree $Div(\mathcal{B}_i)$ based on
            Eq. (24);
        end
        Select $\mathcal{B}^* = \text{argmax}_{\mathcal{B}_i \in \mathbb{U}^*} \mathcal{C}(\mathcal{B}_i)$;
        Query the label of $\mathcal{B}^*$, denoted by $y^*$;
        Let $\mathbb{U} = \mathbb{U} \setminus \mathcal{B}^*$, and $\mathbb{L} = \mathbb{L} \cup (\mathcal{B}^*, y^*)$;
        Update SVM solution $(\mathbf{w}, b)$ based on new $\mathbb{L}$;
    end
end
return $(\mathbf{w}, b)$.
```

FBD will be in the same feature space. Given an unlabeled bag, if the FBD of the bag is large, the dissimilarity degrees of the instances in the bag are also large. This means that the instances are located far away from each other in feature space. In this case, the number of clusters covered by the bag will be large with a high probability, which leads to a large CBD. On the contrary, if the FBD of the bag is small, the CBD will also be small with a high probability. To this end, it can be seen that the CBD and FBD are generally consistent when the same kernel function is adopted.

E. Complexity Analysis

We now give an analysis on the time complexity of selecting one bag in Algorithms 3 and 4. Given an iteration, suppose the numbers of labeled bags, unlabeled bags, selective candidates, features, and instances per bag are l, u, m, d, and n, respectively.

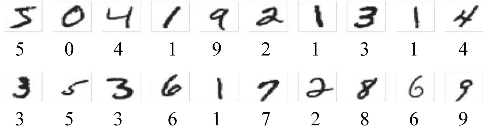


Fig. 5. Training samples in MNIST dataset.

The complexity of Algorithm 3 is focused on the informativeness calculation and kernel k-means algorithm. Making prediction for one testing sample by an SVM has the complexity of $\mathcal{O}(sd)$, where s is the number of support vectors (SVs). We assume that all the $l \times n$ training instances are SVs, thus making predictions for instances in u unlabeled bags and calculating informativeness have the highest complexity of $\mathcal{O}(ln^2du)$. Suppose the number of clustering iteration in Algorithm 2 is ITER, and each cluster has the same number of instances, then the kernel k-means algorithm has the complexity of $\mathcal{O}(\text{ITER} \times mn \times (\frac{mn}{k})^2 \times d)$. Finally, the complexity for selecting one bag in Algorithm 3 is computed as $\mathcal{O}_1 = \mathcal{O}(ln^2du) + \mathcal{O}(\text{ITER} \times mn \times (\frac{mn}{k})^2 \times d)$. If $m \approx k$, then $\mathcal{O}_1 \approx \mathcal{O}(ln^2du) + \mathcal{O}(\text{ITER} \times n^3dm)$

The complexity of Algorithm 4 is focused on the informativeness calculation and dissimilarity degree computation. The informativeness calculation has the same complexity as Algorithm 3, i.e, $\mathcal{O}(ln^2du)$. Computing the dissimilarity degree for one instance based on (23) has the complexity of $\mathcal{O}(nd)$. Thus, the complexity for selecting one bag in Algorithm 3 is computed as $\mathcal{O}_2 = \mathcal{O}(ln^2du) + \mathcal{O}(n^2dm)$. In general, the complexity of Algorithm 4, i.e., MIAL with FBD, is lower than that of Algorithm 3, i.e., MIAL with CBD.

IV. EXPERIMENTAL COMPARISONS

In this section, we will conduct experimental comparisons to show the feasibility and effectiveness of the proposed algorithms.

A. Learning Strategies for Performance Comparison

A total of 13 learning strategies are listed in Table I for performance comparison. Among them, method 1 is a baseline that randomly selects a bag for query during each iteration. Methods

TABLE II
Detailed Description of the MNIST MIL Datasets and Corel MIL Datasets

Datasets	#Features	#Bags		#Instances		#Instances Per Bag			
		positive	negative	positive	negative	min	max	avg	std
Digit “0”	65	100	100	1329	3827	10	40	25.780	8.694
Digit “1”	65	100	100	1298	3620	10	40	24.590	8.934
Digit “2”	65	100	100	1309	3727	10	40	25.180	9.079
Digit “3”	65	100	100	1249	3657	10	40	24.530	9.208
Digit “4”	65	100	100	1311	3708	10	40	25.095	9.209
Digit “5”	65	100	100	1301	3695	10	40	24.980	8.795
Digit “6”	65	100	100	1450	3710	10	40	25.800	8.940
Digit “7”	65	100	100	1330	3818	10	40	25.740	8.929
Digit “8”	65	100	100	1146	3845	10	40	24.955	8.672
Digit “9”	65	100	100	1216	4002	10	40	26.090	9.339
Multi-digit	65	1000	1000	3327	7778	1	10	5.553	2.845
Elephant	230	100	100	762	629	2	13	6.96	2.49
Fox	230	100	100	647	673	2	13	6.60	2.32
Tiger	230	100	100	544	676	1	13	6.10	2.41

2–7 implement Algorithm 1 and query the most valuable bags according to different informativeness measurements, as introduced in Section II-C. The proposed diversity criteria are incorporated with three well-performing informativeness measurements, i.e., minimum instance margin, Noisy-Or model, and Fisher information. As a result, methods 8, 10, and 12 implement Algorithm 3 by incorporating the CBD, and methods 9, 11, and 13 implement Algorithm 4 by incorporating the FBD.

B. Datasets

We conduct experimental comparisons on two groups of datasets, i.e., newly generated MIL datasets from MNIST handwritten digit image recognition problem and existing MIL datasets from Corel content-based image retrieval problem.

1) Handwritten Digit Image Datasets: The MNIST handwritten digit image recognition problem[1] is a task that aims to distinguish 0–9 handwritten digits from approximately 250 writers, as shown in Fig. 5. This problem contains 60 000 training samples and 10 000 testing samples. The raw information of each sample is composed of $28 \times 28 = 784$ gray level pixels with each pixel value $\in \{0, \ldots, 255\}$. We use the gradient-based method[2] presented in [43] and [44] to extract the gradient histogram features, and construct a 2172-dimensional feature vector. Furthermore, in order to generate a compact dataset, we perform a feature selection (FS) process. Since it is time-consuming to conduct FS on a 2172-dimensional feature vector for 60 000 training samples, we divide the features into 22 subsets with the first 21 subsets containing 100 features and the last subset containing 72 features. Then, sequential forward FS (SFFS) [45] is performed on each subset separately. In general, SFFS is a state-of-the-art FS method based on a bottom-up greedy approach. It first initializes an empty feature set $\mathcal{F}_i = \{\emptyset\}, i = 0$; then, it iteratively selects the feature that results in the highest objective function, i.e., $\mathfrak{f}^* = \arg\max_{\mathfrak{f} \in \mathcal{F}_i} [\text{obj}(\mathcal{F}_i \cup \mathfrak{f})]$ and update $\mathcal{F}_{i+1} = \mathcal{F}_i \cup \mathfrak{f}^*, i = i + 1$, until the stopping criterion is satisfied. In this paper, we utilize the function *sequentialfs* in the Statistics and Machine Learning Toolbox of MATLAB with default settings. Finally, the selection results of the 22 subsets are combined and a 65-dimensional feature vector is constructed.

For each class, we generate a single-digit MIL dataset with 100 positive bags and 100 negative bags. We randomly assign the number of instances in a bag from $[10, 40]$. Take digit 0 as an example, and assume the number of instances in the ith bag is n_i. If the bag is negative, the instances are randomly selected from samples of digits 1–9; if the bag is positive, an integer $n_i^+ (1 \leq n_i^+ \leq n_i)$ is first generated as the number of positive instances, then the positive and negative instances are, respectively, selected from samples of digit 0 and digits 1–9 randomly.

In addition, we generate a multidigit MIL dataset with 1000 positive bags and 1000 negative bags. The positive and negative instances are selected from samples of digits 0–4 and 5–9, respectively. The number of instances in a bag is chosen from $[1, 10]$. Finally, the detailed descriptions of the generated MIL datasets are listed in Table II. The task is to identify whether some specific digits exist in a set of handwritten digits.

2) Content-Based Image Retrieval Datasets: The Corel MIL image datasets[3] simulate some content-based image retrieval tasks that aim to distinguish a specific kind of content from other background pictures. Three datasets, i.e., elephant, fox, and tiger, are used. Detailed descriptions of these datasets are listed in Table II. The task is to identify whether a specific animal exists in a set of images.

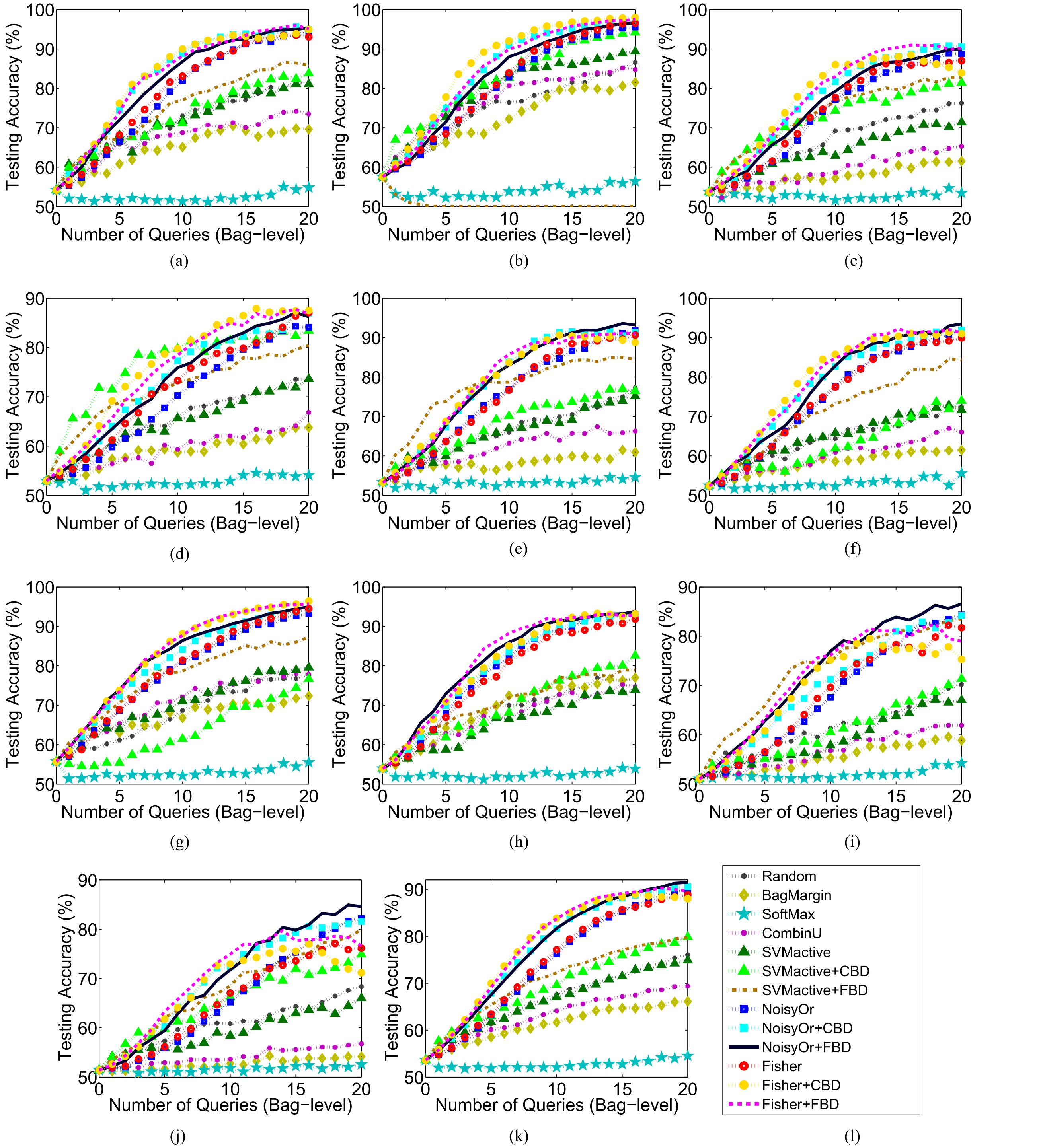
C. Experimental Settings

For the MNIST MIL datasets, 50% bags are randomly selected as the training set, and the remaining 50% bags are taken as the testing set. The learning starts with two positive bags and two negative bags. For the Corel MIL datasets, the number of

[1] http://yann.lecun.com/exdb/mnist

[2] http://www.cs.berkeley.edu/~smaji/projects/digits

[3] http://www.cs.columbia.edu/~andrews/mil/datasets.html



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 6. Performance comparison of different learning strategies on MNIST MIL datasets. (Base-learner: mi-SVM). (a) Digit“0” (50 trials), (b) digit “1” (50 trials), (c) digit “2” (50 trials), (d) digit “3” (50 trials), (e) digit “4” (50 trials), (f) digit “5” (50 trials), (g) digit “6” (50 trials), (h) digit “7” (50 trials), (i) digit “8” (50 trials), (j) digit “9” (50 trials), (k) average result for ten digits, and (l) legend.

instances in the bags is much smaller. Thus, 70% bags are randomly selected as the training set, and the remaining 30% bags are taken as the testing set. The learning starts with ten positive bags and ten negative bags.

In diversity-based strategies, the learner retains $m = 10$ informative unlabeled bags as the candidates during each iteration, and selects the most diverse one to query. Besides, for the kernel k-means clustering algorithm, the number of clusters k is fixed as the average number of instances in the unlabeled bags, and for the softmax model, the parameter α is set as 1. The learning stops after 20 unlabeled bags have been queried or the selective pool becomes empty.

TABLE III
INFORMATIVENESS VALUES OF DIFFERENT UNLABELED BAGS IN A LEARNING ITERATION

Criterion	SVMactive	BagMargin	SoftMax	CombinU	NoisyOr	Fisher
Bag 1	**475.6868**	0.7348	0.6480	0.5555	0.0000	0.0007
Bag 2	4.6145	**0.9586**	0.6864	0.5960	0.0001	0.0138
Bag 3	14.0880	0.6534	**0.6931**	0.5400	0.0000	0.0000
Bag 4	4.6145	0.9586	0.6864	**0.5960**	0.0001	0.0138
Bag 5	10.8069	0.5376	0.5948	0.4941	**0.0820**	3.2402
Bag 6	4.7218	0.5451	0.5841	0.4796	0.0747	**3.3281**

Note: For each criterion, the highest informativeness value is in bold face.

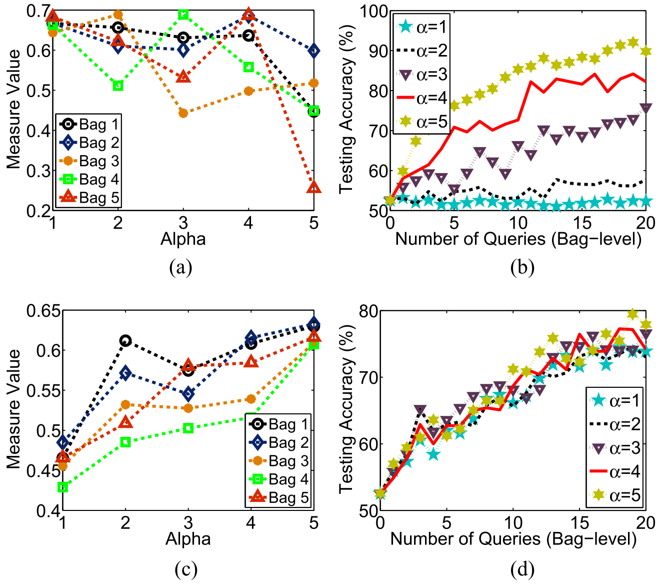


Fig. 7. Sensitivity analysis of parameter α for (a) and (b) SoftMax and (c) and (d) CombinU.

For fair comparison, mi-SVM is employed as the base classifier with parameter $C = 100$, and Gaussian kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$ is adopted with $\sigma = 1$ for all the learning strategies, which also serves as the kernel function in kernel k-means clustering and the similarity relation in fuzzy rough sets. To avoid the random effect, 50 trials are conducted on each dataset and the average results are recorded. The experiments are performed under MATLAB 7.9.0 with the "svmtrain" and "svmpredict" functions of libsvm, which are executed on a computer with a 3.16-GHz Intel Core 2 Duo CPU, a 4-GB memory, and 64-b Windows 7 system.

D. *Result Discussion*

1) Result on MNIST Datasets: Fig. 6 demonstrates the average testing accuracy of 50 trials for each learning strategy on the ten single-digit MNIST MIL datasets. It is observed that the accuracy of mi-SVM trained on the initial training set (i.e., two positive bags and two negative bags) is around 50%, which is just slightly higher than a random guess. By querying new unlabeled bags, the accuracy increases gradually. Basically, we have the following observations.

1) Among the six informativeness measurements (i.e., SVMactive, BagMargin, SoftMax, CombinU, Noisy Or, and Fisher), NoisyOr and Fisher are the best performing ones, which can always achieve higher accuracy than others in the entire learning process; SVMactive is also a competitive method, but its advantage over others is not obvious; BagMargin, SoftMax, and CombinU perform even worse than the baseline Random on most datasets. The reason could be found in Table III, which lists the informativeness values of the selected bags by six measurements in the first learning iteration of *Digit 0*. When considering SVMactive, NoisyOr, or Fisher, the informativeness value of the selected bag is much higher than that of the other bags, however, when considering BagMargin, SoftMax, or CombinU, all the bags have very similar informativeness values. That is to say, the rigorous computation on BagMargin, SoftMax, and CombinU may weaken the differences of the informativeness values among different bags, as a result, the selection may be unreliable due to trivial advantage.
2) The low accuracy of SoftMax and CombinU can be further explained by a sensitivity analysis of parameter α in the softmax function, i.e., (7). Fig. 7 demonstrates the measure values for five unlabeled bags and the learning trends of SoftMax and CombinU with different settings of α on *Digit 0*. For SoftMax, it can be observed that a larger α leads to a higher difference of measure values among the bags, accordingly, the learning performance is improved a lot. This phenomenon also exists for CombinU, but the impact is much smaller. In a word, the performances of SoftMax and CombinU are sensitive to parameter α. In order to get competitive results, we have to make additional efforts for parameter tuning. However, this problem does not exist for other methods. For instance, there is no additional parameter for SVMactive, BagMargin, and NoisyOr, while the only parameter for Fisher is the kernel parameter, which is exactly the same with the one in the SVM. This is also a reason why we did not implement the diversity criteria with SoftMax and CombinU.
3) By incorporating CBD or FBD, the performances of SVMactive, NoisyOr, and Fisher have been improved in most cases. For NoisyOr and Fisher, the improvements achieved by CBD and FBD are similar. This is because the adoption of a fixed kernel function and kernel parameter makes CBD and FBD intrinsically compatible for an informativeness measurement. For SVMactive, FBD outperforms CBD in most cases. This could be due to the fact that SVMactive evaluates an unlabeled bag by its minimum instance margin, i.e., the selection of a bag will only depend on the most valuable instance in it, all the other instances will be useless. As analyzed in Sections III-B and III-C, CBD is based on a clustering process of all the instances, whereas FBD is directly related to the most valuable instance with regard to the evaluation targets. As a result, FBD is more suitable to be incorporated

TABLE IV
AVERAGE TESTING ACCURACY (%) OF THE 20 LEARNING ITERATIONS (UPPER RESULT) AND FINAL TESTING ACCURACY (%) AFTER THE LEARNING PROCESS IS FINISHED (LOWER RESULT)

Datasets	Random	BagMargin	SoftMax	CombinU	▲	▲+CBD	▲+FBD	⋇	⋇+CBD	⋇+FBD	★	★+CBD	★+FBD
Digit “0”	71.49	64.56	52.41	67.21	71.26	72.30↑	74.84↑	78.12	82.93↑	81.52↑	78.88	**83.00**↑	82.66↑
	81.08	69.60	54.84	73.50	81.08	83.80↑	85.84↑	94.34	94.88↑	**95.30**↑	93.04	94.90↑	95.10↑
Digit “1”	74.44	72.22	54.02	76.84	79.21	81.82↑	50.60⇊	80.10	84.14↑	82.46↑	80.83	**86.12**↑	84.21↑
	86.54	81.54	56.40	84.74	89.36	94.18↑	50.10⇊	95.52	97.04↑	96.62↑	96.52	**98.00**↑	97.56↑
Digit “2”	66.57	57.60	52.74	59.40	64.03	72.31↑	73.34↑	73.47	77.39↑	75.78↑	73.97	77.72↑	**78.38**↑
	76.24	61.54	53.46	65.30	71.38	81.42↑	82.74↑	88.74	**90.60**↑	89.84↑	87.00	83.90⇊	90.20↑
Digit “3”	65.03	58.76	52.77	59.49	64.47	76.01↑	71.24↑	69.31	73.04↑	72.83↑	71.35	**76.06**↑	75.00↑
	73.52	63.76	54.10	66.84	73.64	83.40↑	80.42↑	84.10	86.92↑	86.22↑	87.06	**87.50**↑	86.86⇊
Digit “4”	65.98	57.96	53.16	61.77	65.58	68.05↑	76.88↑	74.94	**78.74**↑	78.66↑	74.91	78.54↑	78.73↑
	77.30	60.96	54.60	66.30	75.20	76.46↑	84.54↑	91.86	91.36⇊	**93.24**↑	90.68	88.78⇊	91.04↑
Digit “5”	64.14	58.14	52.95	59.93	64.81	62.83⇊	71.08↑	74.65	77.73↑	77.68↑	74.52	**79.37**↑	79.04↑
	72.48	61.50	55.60	66.06	71.62	73.98↑	84.34↑	91.80	91.96↑	**93.46**↑	89.96	91.00↑	91.24↑
Digit “6”	67.89	66.66	52.97	70.33	69.93	63.16⇊	76.58↑	78.11	80.69↑	81.83↑	78.61	82.85↑	**82.92**↑
	77.68	72.38	55.52	77.88	79.56	76.68⇊	87.28↑	93.20	94.46↑	94.90↑	94.46	**96.40**↑	95.70↑
Digit “7”	68.41	68.51	52.37	66.73	65.24	69.56↑	70.24↑	78.41	79.14↑	80.90↑	77.10	79.97↑	**81.19**↑
	76.74	76.98	53.90	74.66	73.90	82.62↑	78.88↑	92.92	92.68⇊	**93.80**↑	91.84	93.18↑	92.76↑
Digit “8”	61.05	55.27	51.95	56.71	59.00	60.75↑	72.88↑	67.71	69.84↑	**72.67**↑	67.50	69.56↑	71.64↑
	70.22	58.86	54.26	61.94	67.02	71.36↑	83.56↑	84.34	84.20⇊	**86.58**↑	81.72	75.34⇊	79.00⇊
Digit “9”	60.74	52.68	51.63	54.11	59.08	65.13↑	67.52↑	65.83	69.90↑	**70.11**↑	65.18	67.84↑	70.01↑
	68.36	54.18	52.56	56.76	66.02	74.90↑	79.96↑	82.18	81.56⇊	**84.60**↑	76.16	71.20⇊	76.64↑
Avg.	66.57	61.24	52.70	63.25	66.26	69.19↑	70.52↑	74.06	77.36↑	77.44↑	74.29	78.10↑	**78.38**↑
	76.02	66.13	54.52	69.40	74.88	79.88↑	79.77↑	89.90	90.57↑	**91.46**↑	88.84	88.02⇊	89.61↑

Note: Due to space limit, we denote “SVMactive,” “NoisyOr,” and “Fisher” as “▲”, “⋇,” and “★,” respectively. For each dataset, the highest average accuracy is in bold face. The symbols of ↑ and ⇊, respectively, represent that the average accuracy is improved or not by incorporating CBD or FBD.

with SVMactive. However, an exceptional case exists in Fig. 6(b), where SVMactive + FBD demonstrates a decreasing learning trend on *Digit 1*. This could be caused by a bad selection in the first learning iteration, which obstructs the classifier to converge to the optimal one at the beginning.

Furthermore, Fig. 6(k) shows the average performance of the learning strategies on the ten single-digit MNIST MIL datasets. Overall speaking, CBD and FBD can achieve very similar improvements for the informativeness measurements.

Table IV reports the mean accuracy of the 20 learning iterations and the final accuracy after the learning stops, respectively. For each dataset, the best results are highlighted in bold face. It is observed that the best results are always achieved by incorporating a diversity criterion. Among them, Fisher + FBD and NoisyOr+FBD give the best average results for mean accuracy and final accuracy, respectively. Besides, we use ↑ and ⇊ to demonstrate whether the diversity criteria can improve the performance of its single informativeness-based strategy. Obviously, accuracy improvement is achieved on most datasets for both CBD and FBD with regard to all the three informativeness measurement, i.e., SVMactive, NoisyOr, and Fisher.

Table V reports the average time cost during each learning iteration and the average number of iterations for training mi-SVM by different strategies. The time cost during each iteration mainly consists of two parts: training base classifiers and evaluating unlabeled bags. The first part is directly related to the number of training iterations for mi-SVM. The selection of a valuable bag can not only improve the learning performance, but also reduce the number of training iterations and force the SVM to converge to the optimal one faster. It is observed that the average number of training iterations for mi-SVM in a diversity-based strategy (e.g., Fisher + CBD or Fisher + FBD) is very close to that in its single informativeness-based strategy (e.g., Fisher), which demonstrates that the incorporation of diversity to informativeness will not increase the training complexity. As for the bag evaluation part, CBD has a higher complexity than FBD, which is clear from Table V. However, in real-world AL applications, labeling a sample usually takes much more time than selecting a sample. For instance, it may take several seconds to several minutes for labeling a sample, whereas the selecting part just takes milliseconds. Thus, the time complexity of all the strategies is in an acceptable range.

We also make some statistical tests on the results listed in Table IV. Paired Wilcoxon's signed rank test is performed, which is a famous nonparametric statistical hypothesis test for assessing whether there exists significant difference between two sets of results. The corresponding p-values are reported in Table VI, and the significance level 0.05 is adopted. If the p-value is smaller than 0.05, the two referred methods are considered as statistically different. It can be seen that almost all the diversity-based strategies are statistically different from the single informativeness-based strategies by considering both mean accuracy and final accuracy. However, the diversity-based strategies may have no essential difference from each other in some cases. This is consistent with the results shown in Fig. 6, where the learning trends of some diversity-based strategies (e.g., Fisher + CBD and Fisher + FBD) are similar during the entire process.

Finally, Fig. 8 demonstrates the AL performance on the multi-digit MNIST MIL dataset. According to the above-presented re-

TABLE V
AVERAGE TIME (SECONDS) FOR SELECTING ONE BAG AND AVERAGE NUMBER OF ITERATIONS FOR TRAINING MI-SVM

Data sets	Random	BagMargin	SoftMax	CombinU	▲	▲+CBD	▲+FBD	※	※+CBD	※+FBD	★	★+CBD	★+FBD
Digit "0"	0.1976	0.2184	0.6125	0.2212	0.2659	1.2487	0.3766	0.0914	1.0056	0.1720	0.1629	1.0290	0.1984
	(4.2)	(4.5)	(7.5)	(4.5)	(4.6)	(5.5)	(4.5)	(2.6)	(2.4)	(2.6)	(2.6)	(2.4)	(2.5)
Digit "1"	0.1654	0.1445	0.5238	0.1363	0.1773	1.0169	0.3410	0.0696	0.9162	0.1563	0.1415	0.9288	0.1892
	(3.8)	(3.8)	(7.4)	(3.4)	(3.5)	(4.1)	(2.6)	(2.4)	(2.3)	(2.5)	(2.3)	(2.2)	(2.4)
Digit "2"	0.2305	0.2629	0.5544	0.2700	0.3410	1.1897	0.4531	0.0990	0.9653	0.1934	0.1806	0.9841	0.2221
	(4.8)	(4.9)	(7.3)	(5.0)	(5.3)	(5.3)	(5.1)	(2.9)	(2.7)	(2.9)	(2.8)	(2.7)	(2.8)
Digit "3"	0.2462	0.2296	0.5343	0.2414	0.3167	1.1243	0.4283	0.0950	0.9460	0.1799	0.1690	0.9650	0.2067
	(5.0)	(4.8)	(7.0)	(5.0)	(5.1)	(5.0)	(4.9)	(2.9)	(2.8)	(2.9)	(2.7)	(2.7)	(2.7)
Digit "4"	0.2401	0.2519	0.5043	0.2493	0.3094	1.2057	0.3615	0.0966	0.9867	0.1892	0.1813	1.0021	0.2157
	(4.9)	(5.0)	(7.3)	(5.0)	(4.9)	(5.5)	(4.1)	(2.7)	(2.7)	(2.8)	(2.6)	(2.6)	(2.6)
Digit "5"	0.2597	0.2495	0.5652	0.2556	0.3431	1.3694	0.4189	0.0912	0.9595	0.1818	0.1722	0.9771	0.2071
	(5.2)	(5.0)	(7.3)	(5.0)	(5.3)	(6.9)	(4.9)	(2.7)	(2.6)	(2.8)	(2.7)	(2.6)	(2.6)
Digit "6"	0.2297	0.2205	0.6099	0.1981	0.2861	1.3965	0.3963	0.0801	1.0177	0.1614	0.1603	1.0307	0.1914
	(4.7)	(4.5)	(7.7)	(4.1)	(4.7)	(6.5)	(4.7)	(2.5)	(2.5)	(2.5)	(2.5)	(2.4)	(2.5)
Digit "7"	0.2042	0.1845	0.6318	0.2044	0.3053	1.2729	0.4380	0.0846	1.0221	0.1652	0.1676	1.0480	0.2019
	(4.3)	(4.0)	(7.5)	(4.3)	(4.9)	(5.6)	(5.0)	(2.6)	(2.5)	(2.5)	(2.5)	(2.5)	(2.5)
Digit "8"	0.3123	0.2973	0.5615	0.2769	0.4077	1.4177	0.4557	0.1141	0.9954	0.2079	0.1950	1.0199	0.2281
	(5.6)	(5.3)	(7.2)	(5.0)	(5.9)	(6.9)	(5.1)	(3.2)	(3.2)	(3.2)	(3.0)	(3.1)	(3.0)
Digit "9"	0.3144	0.3503	0.6202	0.3578	0.4208	1.3723	0.4564	0.1149	1.0873	0.2160	0.2000	1.1141	0.2509
	(5.4)	(5.8)	(7.4)	(5.7)	(5.7)	(6.0)	(4.9)	(3.1)	(3.0)	(3.1)	(2.9)	(2.9)	(2.9)
Avg.	0.2400	0.2409	0.5718	0.2411	0.3173	1.2614	0.4126	0.0936	0.9902	0.1823	0.1731	1.0099	0.2111
	(4.8)	(4.8)	(7.4)	(4.7)	(5.0)	(5.7)	(4.6)	(2.8)	(2.7)	(2.8)	(2.6)	(2.6)	(2.7)

Note: Due to space limit, we denote "SVMactive", "NoisyOr" and "Fisher" as "▲", "※" and "★" respectively.

TABLE VI
PAIRED WILCOXON'S SIGNED-RANK TESTS (p-VALUES)

Method	BagMargin	SoftMax	CombinU	▲	▲+CBD	▲+FBD	※	※+CBD	※+FBD	★	★+CBD	★+FBD
Random	0.0039†	0.0020†	0.0195†	0.4922	0.1309	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
	0.0039†	0.0020†	0.0039†	0.1641	0.0098†	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
BagMargin	–	0.0020†	0.0137†	0.0059†	0.0059†	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
	–	0.0020†	0.0039†	0.0039†	0.0020†	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
SoftMax	–	–	0.0020†	0.0020†	0.0020†	0.0039†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
	–	–	0.0020†	0.0020†	0.0020†	0.0039†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
CombinU	–	–	–	0.0098†	0.0371†	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
	–	–	–	0.0039†	0.0039†	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
▲	–	–	–	–	0.1055	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
	–	–	–	–	0.0137†	0.0840	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†	0.0020†
▲+CBD	–	–	–	–	–	0.2324	0.0488†	0.0059†	0.0059†	0.0273†	0.0020†	0.0039†
	–	–	–	–	–	0.3223	0.0020†	0.0020†	0.0020†	0.0020†	0.0059†	0.0020†
▲+FBD	–	–	–	–	–	–	0.4922	0.0137†	0.0039†	0.2754	0.0098†	0.0039†
	–	–	–	–	–	–	0.0020†	0.0020†	0.0020†	0.0098†	0.1055	0.0098†
※	–	–	–	–	–	–	–	0.0020†	0.0020†	0.5566	0.0020†	0.0020†
	–	–	–	–	–	–	–	0.1934	0.0020†	0.1934	0.4922	0.9219
※+CBD	–	–	–	–	–	–	–	–	1.0000	0.0020†	0.1602	0.0195†
	–	–	–	–	–	–	–	–	0.0898	0.0078†	0.2871	0.4316
※+FBD	–	–	–	–	–	–	–	–	–	0.0020†	0.3750	0.0273†
	–	–	–	–	–	–	–	–	–	0.0098†	0.1309	0.1934
★	–	–	–	–	–	–	–	–	–	–	0.0020†	0.0020†
	–	–	–	–	–	–	–	–	–	–	0.6250	0.0840
★+CBD	–	–	–	–	–	–	–	–	–	–	–	0.5566
	–	–	–	–	–	–	–	–	–	–	–	0.3750

Note: In each comparison, the upper and lower results are, respectively, the p-values of the Wilcoxon's signed rank tests on the mean accuracy and final accuracy in Table IV. For each test, † represents that the two referred methods are significantly different with the significance level 0.05.

sults, Fisher + CBD and Fisher + FBD have achieved the best performance. Thus, we only implement Fisher, Fisher + CBD, Fisher + FBD, and the baseline Random on this dataset. Obviously, the performance of Fisher has been improved, especially by Fisher + FBD.

2) Result on Corel Data sets: For simplicity, we only implement Random, NoisyOr, NoisyOr + CBD, NoisyOr+FBD, Fisher, Fisher + CBD, and Fisher + FBD for this group of datasets. Fig. 9 demonstrates the average testing accuracy of 50 trials for each learning strategy. Unfortunately, the diversity criteria (i.e., CBD and FBD) fail to improve the performance of the single informativeness-based strategies (i.e,. NoisyOr and Fisher) on these tasks. The underlying reason could be found in Fig. 10, which demonstrates the instance distribution of different unlabeled bags in the first two feature dimensions of the MNIST datasets and Corel datasets. It is clear that the MNIST datasets possess the two basic characteristics described in Section III-A, i.e., the number of instances in different bags differs a lot and the instance distribution is highly irregular. However, these two characteristics are not obvious in the Corel

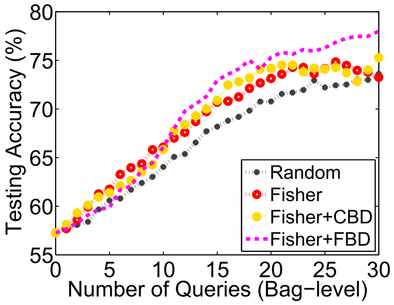
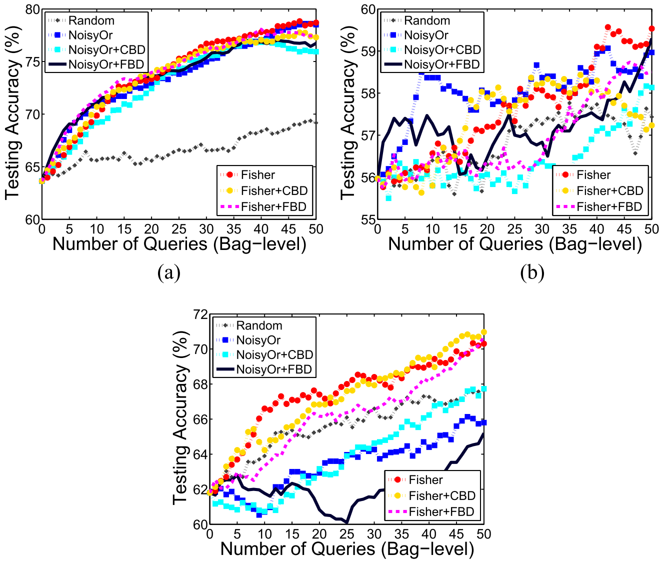


Fig. 8. Performance comparison on the multidigit MIL dataset. (Base-learner: mi-SVM.)

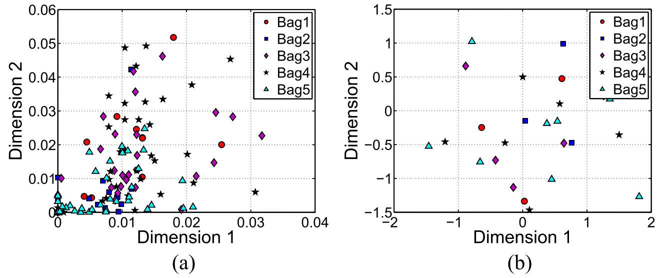


(a)

(b)

(c)

Fig. 9. Performance comparison of different learning strategies on Corel MIL datasets. (Base-learner: mi-SVM). (a) Elephant (50 trials). (b) Fox (50 trials). (c) Tiger (50 trials).



(a)

(b)

Fig. 10. Different bags in MNIST and Corel datasets. (a) Bags in MNIST datasets. (b) Bags in Corel datasets.

datasets. As a result, the performance may even be decreased if the data do not satisfy the premises described in Section III-A. Furthermore, it can be seen from Table II that the number of instances in a bag for the Corel datasets is much smaller than that of the MNIST datasets, and the dimensionality of the feature vector for the Corel datasets is much higher than that of the MNIST datasets. In such a high-dimensional feature space, the distribution for a few instances will be very sparse. That is to say, all the instances are far away from each other and the diversity evaluation is not a necessary step.

Algorithm 5: Pseudo-Code for mi-SVM Optimization Heuristics

```
Input:
    Multiple-instance training set $\mathbb{S} = \{(\mathcal{B}_i, y_i)\}_{i=1}^n$.
Output:
    SVM solution $(\mathbf{w}, b)$.
Initialize $y_{ij} = y_i$ for each $\mathcal{B}_{ij} \in \mathcal{B}_i$;
repeat
    Compute SVM solution $(\mathbf{w}, b)$ with imputed labels $y_{ij}$;
    for each positive bag $\mathcal{B}_i$ do
        Compute $f_{ij} = \mathbf{w}^{\mathrm{T}}\mathcal{B}_{ij} + b$ for each $\mathcal{B}_{ij} \in \mathcal{B}_i$;
        Set $y_{ij} = \mathrm{sign}(f_{ij})$ for each $\mathcal{B}_{ij} \in \mathcal{B}_i$;
        if $\sum_{\mathcal{B}_{ij}\in\mathcal{B}_i}(1 + y_{ij})/2 == 0$ then
            Compute $j^* = \mathrm{argmax}_j\, f_{ij}$;
            Set $y_{ij^*} = 1$;
        end
    end
until imputed labels $y_{ij}$ have no change;
return $(\mathbf{w}, b)$.
```

Algorithm 6: Pseudo-Code for MI-SVM Optimization Heuristics

```
Input:
    Multiple-instance training set $\mathbb{S} = \{(\mathcal{B}_i, y_i)\}_{i=1}^n$.
Output:
    SVM solution $(\mathbf{w}, b)$.
Initialize $\mathbf{x}_i = \sum_{\mathcal{B}_{ij}\in\mathcal{B}_i}\mathcal{B}_{ij}/|\mathcal{B}_i|$ for each positive bag $\mathcal{B}_i$;
repeat
    Compute SVM solution $(\mathbf{w}, b)$ with all instances in
    negative bags and positive examples $\{\mathbf{x}_i : y_i = 1\}$;
    for each positive bag $\mathcal{B}_i$ do
        Compute $f_{ij} = \mathbf{w}^{\mathrm{T}}\mathcal{B}_{ij} + b$ for each $\mathcal{B}_{ij} \in \mathcal{B}_i$;
        Compute $j^* = \mathrm{argmax}_j\, f_{ij}$;
        Set $\mathbf{x}_i = \mathcal{B}_{ij^*}$;
    end
until selector variables $j^*$ have no change;
return $(\mathbf{w}, b)$.
```

V. Conclusion

In this paper, two diversity criteria have been proposed to evaluate unlabeled bags in SVM-based MIAL, i.e,. CBD and FBD, which measure the diversity of an unlabeled bag by the kernel k-means clustering algorithm and lower approximations in fuzzy rough sets. By incorporating CBD and FBD with traditional informativeness measurements, the learner can query the unlabeled bag with both high informativeness and diversity. Moreover, the kernel function adopted in the SVM also serves as the kernel function in kernel k-means clustering and the similarity relation in fuzzy rough sets, which makes CBD and FBD intrinsically compatible with SVM. Experimental comparisons demonstrate that the diversity criteria are effective to improve performance when the number of instances in different bags differs a lot and the instance distribution is highly irregular.

APPENDIX

The heuristic optimization models of mi-SVM and MI-SVM are presented as Algorithms 5 and 6, respectively.

REFERENCES

[1] P. Auer, "On learning from multi-instance examples: Empirical evaluation of a theoretical approach," in *Proc. Int. Conf. Mach. Learn.*, 1997, vol. 97, pp. 21–29.

[2] T. Dietterich, R. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1/2, pp. 31–71, 1997.

[3] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *Proceedings 17th International Conference on Machine Learning.* San Mateo, CA, USA: Morgan Kaufmann, 2000, pp. 1119–1125.

[4] 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.

[5] D. Wang, J. Li, and B. Zhang, "Multiple-instance learning via random walk," in *Proc. Eur. Conf. Mach. Learn.*, 2006, pp. 473–484.

[6] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1249–1256.

[7] A. Zafra and S. Ventura, "G3P-MI: A genetic programming algorithm for multiple instance learning," *Inf. Sci.*, vol. 180, no. 23, pp. 4496–4513, 2010.

[8] O. Kundakcioglu, O. Seref, and P. Pardalos, "Multiple instance learning via margin maximization," *Appl. Numer. Math.*, vol. 60, no. 4, pp. 358–369, 2010.

[9] J. Bolton, P. Gader, H. Frigui, and P. Torrione, "Random set framework for multiple instance learning," *Inf. Sci.*, vol. 181, pp. 2061–2070, 2011.

[10] 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.

[11] C. Jiao and A. Zare, "Functions of multiple instances for learning target signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4670–4686, Aug. 2015.

[12] D. Li, J. Peng, Z. Li, and Q. Bu, "LSA based multi-instance learning algorithm for image retrieval," *Signal Process.*, vol. 91, pp. 1993–2000, 2011.

[13] T.-C. Lin, M.-C. Yang, C.-Y. Tsai, and Y.-C. F. Wang, "Query-adaptive multiple instance learning for video instance retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1330–1340, Apr. 2015.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 2000.

[15] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, vol. 15, pp. 561–568.

[16] M. I. Mandel and D. P. Ellis, "Multiple-instance learning for music information retrieval," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, 2008, pp. 577–582.

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

[18] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2057–2063.

[19] C. Zeng, H. Ma, and A. Ming, "Fast human detection using mi-sVM and a cascade of HOG-LBP features," in *Proc. 17th IEEE Int. Conf. Image Process.*, 2010, pp. 3845–3848.

[20] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.

[21] R. Wang, S. Kwong, and D. Chen, "Inconsistency-based active learning for support vector machines," *Pattern Recognit.*, vol. 45, no. 10, pp. 3751–3767, 2012.

[22] R. Wang, D. Chen, and S. Kwong, "Fuzzy-rough-set-based active learning," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 6, pp. 1699–1704, Dec. 2014.

[23] R. Wang, C.-Y. Chow, and S. Kwong, "Ambiguity based multiclass active learning," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 1, pp. 242–248, Feb. 2016.

[24] B. Zhang, Y. Wang, and F. Chen, "Multilabel image classification via high-order label correlation driven active learning," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1430–1441, Mar. 2014.

[25] B. Long, J. Bian, O. Chapelle, Y. Zhang, Y. Inagaki, and Y. Chang, "Active learning for ranking through expected loss optimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1180–1191, May 2015.

[26] L. Chen, X. Tian, and L. Cai, "FIM-based pairwise selection for active learning on imbalanced datasets," in *Proc. IEEE Int. Conf. Syst. Man, Cybern.*, 2015, pp. 1876–1881.

[27] J. Zhang, X. Wu, and V. S. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1095–1107, May 2015.

[28] D. Liu, X. Hua, L. Yang, and H. Zhang, "Multiple-instance active learning for image categorization," in *Advances in Multimedia Modeling: International Multimedia Modeling Conference.* New York, NY USA: Springer. 2009, pp. 239–249.

[29] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2008, pp. 1289–1296.

[30] J. Fu and J. Yin, "Bag-level active multi-instance learning," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery*, 2011, vol. 2, pp. 1307–1311.

[31] S. Srinivas, "A generalization of the noisy-or model," in *Proc. 9th Conf. Uncertainty. Artif. Intell.*, 1993, pp. 208–215.

[32] D. Zhang, F. Wang, Z. Shi, and C. Zhang, "Interactive localized content based image retrieval with multiple-instance active learning," *Pattern Recognit.*, vol. 43, no. 2, pp. 478–484, 2010.

[33] R. Wang and S. Kwong, "Active learning with multi-criteria decision making systems," *Pattern Recognit.*, vol. 47, no. 9, pp. 3106–3119, 2014.

[34] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 417–424.

[35] T. Zhang and F. Oles, "A probability analysis on the value of unlabeled data for classification problems," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 1191–1198.

[36] S. Hoi, R. Jin, and M. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1233–1248, Sep. 2009.

[37] J. Chiang and S. Cheng, "Multiple-instance content-based image retrieval employing isometric embedded similarity measure," *Pattern Recognit.*, vol. 42, no. 1, pp. 158–166, 2009.

[38] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," *Intell. Decis. Support*, vol. 11, pp. 203–232, 1992.

[39] Q. Hu, L. Zhang, S. An, D. Zhang, and D. Yu, "On robust fuzzy rough set models," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 4, pp. 636–651, Aug. 2012.

[40] A. M. Radzikowska and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets Syst.*, vol. 126, no. 2, pp. 137–155, 2002.

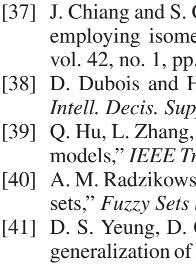
[41] D. S. Yeung, D. Chen, E. C. C. Tsang, J. Lee, and X. Z. Wang, "On the generalization of fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 3, pp. 343–361, Jun. 2005.

[42] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, "Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications," *Int. J. Approx. Reason.*, vol. 51, no. 4, pp. 453–471, 2010.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[44] S. Maji and J. Malik, "Fast and accurate digit classification," Dept. Elect. Eng. Comput. Sci., , Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2009-159, 2009.

[45] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.



Ran Wang (S'09–M'14) received the B.Eng. degree in computer science from the College of Information Science and Technology, Beijing Forestry University, Beijing, China, in 2009, and the Ph.D. degree in computer science from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2014.

From 2014 to 2016, she was a Postdoctoral Researcher in the Department of Computer Science, City University of Hong Kong. She is currently an Assistant Professor in the College of Mathematics and Statistics, Shenzhen University, Shenzhen, China. Her current research interests include pattern recognition, machine learning, fuzzy sets and fuzzy logic, and their related applications.

Xi-Zhao Wang (M'03–SM'04–F'12) received the Doctoral degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1998.

From 2001 to 2014, he has been a Full Professor and the Dean of the College of Mathematics and Computer Science, Hebei University, Hebei, China. From 1998 to 2001, he was a Research Fellow in the Department of Computing, Hong Kong Polytechnic University, Hong Kong. Since 2014, he has been a Full Professor in the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. His main research interests include supervised and unsupervised learning, active learning, reinforcement learning, manifold learning, transfer learning, unstructured learning, uncertainty, fuzzy sets and systems, fuzzy measures and integrals, rough set, and learning from big data.

Dr. Wang is a member of the Board of Governors of the IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2005, 2007–2009, 2012–2014), the Chair of the Technical Committee on Computational Intelligence of the IEEE SMC, and a Distinguished Lecturer of the IEEE SMC. He was the Program Co-Chair of the IEEE SMC 2009 and 2010. He has received many awards from the IEEE SMC Society. He is the Editor-in-Chief of the *International Journal of Machine Learning and Cybernetics*. He is also an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, the *Information Sciences Journal*, and the *International Journal of Pattern Recognition and Artificial Intelligence*.

Sam Kwong (M'93–SM'04–F'13) received the B.Sc. degree in electrical engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 1983, the M.S. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree in computer science from the University of Hagen, Hagen, Germany, in 1996.

From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined the Bell Northern Research, Canada as a member of scientific staff. In 1990, he became a Lecturer in the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he is currently a Professor and the Head of the Department of Computer Science. His main research interests include evolutionary computation, video coding, pattern recognition, and machine learning.

Dr. Kwong is an Associate Editor of the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and the *Information Sciences Journal.*

Chen Xu received the B.Sc. and M.Sc. degrees from Xidian University, Xi'an, China, in 1986 and 1989, respectively, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 1992, all in mathematics.

He joined Shenzhen University, Shenzhen, China, in 1992, where he is currently a Professor. From 1999 to 2000, he was a Research Fellow with Kansai University, Suita, Japan, and the University of Hawaii, Honolulu, HI, USA, from 2002 to 2003. His current research interests include image processing, intelligent computing, and wavelet analysis.