IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA ANALYTICS FOR LARGE-SCALE COMPLEX DATA ENVIRONMENTS

Received February 19, 2018, accepted March 15, 2018, date of publication March 21, 2018, date of current version May 9, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2817845*

# Multi-Class Active Learning by Integrating Uncertainty and Diversity

**ZENGMAO WANG[1,2], XI FANG[2], XINYAO TANG[2], AND CHEN WU[1,2], (Member, IEEE)**
[1]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430072, China
[2]School of Computer Science, Wuhan University, Wuhan 430072, China

Corresponding author: Chen Wu (chen.wu@whu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601333 and Grant U1536204, in part by the China Post-Doctoral Science Foundation under Grant 2016T90733, and in part by the Open Research Fund of Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, under Grant 2017LDE003.

**ABSTRACT** Active learning is a promising way to reduce the labeling cost with a limited training samples initially, and then iteratively select the most valuable samples from a large number of unlabeled data for labeling in order to construct a powerful classifier. The goal of active learning is to make the labeled data set has no redundancy as much as possible. Uncertainty and diversity are two important criteria for active learning. Currently, a promising way by combining uncertainty and diversity for active learning is developed. However, many of these methods are designed based on the binary class or uncertainty followed by diversity strategy. They are hard to select the most valuable samples for multiple classes with binary setting with diversity and uncertainty simultaneously. In this paper, we integrate uncertainty and diversity into one formula by multi-class settings. Uncertainty is measured by the margin minimum while diversity is measured by the maximum mean discrepancy, which is popular to measure the distribution between two data sets. By minimizing the upper bound for the true risk of the integrating formula, we find the samples that not only uncertainty but also diversity with each other. We conduct our experiments on 12 benchmark UC Irvine data sets, and the experimental results demonstrate that the proposed method performs better than some other state-of-the-art methods.

**INDEX TERMS** Active learning, multi-class, informative and representative criterion, empirical risk minimization principle.

## I. INTRODUCTION

In many real problems, unlabeled data is easy to obtain, while labeling is usually expensive and time-consuming due to the involvement of human experts. Hence, it is significance to train a good classifier to deal with limited labeled data. Active learning addresses this challenge by querying the most informative samples in order to allow the large amount of unlabeled data to be annotated automatically.

The key component of active learning methods depends on the design of an available criterion for iteratively selecting the most valuable samples. Two primary strategies, i.e. uncertainty and diversity, are widely used to design practical active learning algorithms [1]. Uncertainty measures whether the unlabeled data to improve the generalization ability of the current classifier, which could rapidly shrink all the classifiers. The diversity of the queried samples denotes whether the redundancy exits in the labeled data. Most active learning algorithms simply employ one of the two strategies. The most useful approaches of querying the most informative samples for active learning include expected error reduction [2], [3], query by committee [4]–[6], and the most uncertain criteria [1], [7]–[9]. The disadvantages of such methods include that the queried samples cannot make sure they are different from each other without redundancy. Hence, when training the classifier just with the uncertain samples, this can lead to serious sample bias and consequently undesirable performance [10], [11]. The second active learning strategy aims to query the representative samples for the overall patterns of unlabeled data, which is a famous diversity strategy in current researches [1], [12], [13]. They exploit the structure of unlabeled data and guarantee that it is independent of and identically distributed from the original data. Such methods can perform better when there is few or no initial labeled data. However, since there is a lack of uncertain information,

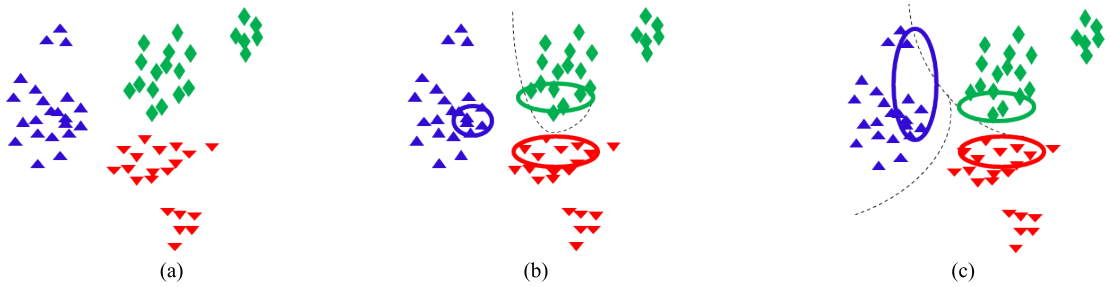



**FIGURE 1.** An illustrative example for selecting most informative samples with uncertainty and diversity (representativeness) for binary and multiple class settings. (a) Data distribution. (b) Samples queried based on binary class setting with uncertainty and diversity. (c) Samples queried based on multiple class setting with uncertainty and diversity.

these methods must query quite a large amount of unlabeled data before the optimal classifier is trained. Furthermore, efficiency also will decline alongside an increase in the queried data.

Deploying either strategy will significantly limit performance. Several researchers have attempted to query samples with high uncertainty and high representativeness (diversity) [14]–[18]. They are mostly ad hoc in terms of measuring the informativeness and representativeness of one sample, leading to suboptimal performance. Recently, Li and Guo [19] attempted to apply both the uncertainty and diversity approaches. They calculated the two terms respectively and then combined the two terms with multiplication and balanced them by a weight. They used conditional entropy as the measure of informativeness and the Gaussian Process framework to measure the representativeness of unlabeled data. In this process, an inverse covariance matrix whose size is equal to the length of the unlabeled data should be calculated. However, if the unlabeled data is too limited, the covariance matrix will be singular or near singular. If the unlabeled data is too large, the computational complexity will increase rapidly, especially if the unlabeled data does not satisfy the requirement of normal or near normal distribution, thereby failing to achieve good performance. Huang *et al*. [11] attempted to use both the uncertainty and diversity approaches in one optimization formulation based on the max-min view [24]. They used unlabeled data in a semi-supervised learning setting in order to boost performance. However, using this method, the queried samples may not preserve the original data distribution. If the data structure does not satisfy semi-supervised assumptions [21]–[23], the performance would not be good enough. In addition, none of the methods described above consider the label information and are designed based on binary class.

Here, we aim to make full use of the label information in order to boost active learning performance. Inspired by [11] and [20], we present a novel active learning approach by integrating uncertainty and diversity for multi-class setting. For diversity strategy, the representativeness is adopted. Hence, we denote the proposed method as Informative and Representative Active learning (IR-AL) for multi-class setting. It aims to address the following two objectives: 1) to query the instances containing the most informative information; and 2) to enforce the query instances that are diverse with each other in the whole active learning procedure. To understand intuitively, we show the binary setting and multiple class setting in Figure.1. Our main contributions to the field include:

1) For the multiple classes, it is anticipated that the label information for each instance will be fully used. Generally, for single label classification, for each instance, the label that will be used in the method is simply either positive or negative, and uncertainty is measured just by one binary classifier. In the proposed IR-AL, for an instance, the label is a vector which just contains one positive value corresponding to the class to which the instance belongs. In this situation, the query instance can be closest to all the classification hyperplanes between the binary classes.

2) We minimize empirical risk both of the uncertainty and diversity simultaneously, and it can make sure the queried samples adjust the uncertain information and diversity information adaptively. Meanwhile, in the proposed formula, a non-convex problem exists. An algorithm is developed to solve it.

3) The proposed method can experience the issue of limited initial labeled data, which might result in a slow convergence in active learning. We adopt the maximum mean discrepancy (MMD) to select the most representative samples as diversity in order to mitigate the lack of uncertain information, and thus to ensure that the method's performance is maximized.

We have demonstrated the proposed algorithm on several UC Irvine (UCI) benchmark datasets. The results indicate that our proposed algorithm is superior to current state-of-the-art active learning methods.

The rest of this paper is organized as follows. Section II will provide a background of the current knowledge on active learning and correlation theory which is relevant to the proposed method. In section III, we will introduce our proposed method in detail, and then describe the experiments in order to demonstrate the efficiency of our method in section IV. Finally, a simple summary about our method and a brief discussion of future research directions will be provided.

## II. PRELIMINARY

In the proposed framework, we adopt the representativeness as diversity strategy. Then we derive an empirical risk for the active learning risk and use the MMD to measure the representative information. Minimizing empirical risk has been successfully applied in machine learning and data mining methods [11], [25], [26]. In [26] and [27], it was demonstrated that minimizing true risk in the context of unseen data distribution is approximated by the summation of empirical risk on the labeled data and a properly designed regularization term, which constrains the complexity of the candidate classifiers.

### A. ACTIVE LEARNING

Active learning can be modeled as a quintuple (T, F, U, Q, S). It is an effective approach for resolving the small sample problem. In active learning, T is the labeled dataset with the limited samples, F is the classifier model trained by T, U is the pool of samples that contains abundant unlabeled samples, Q is a dataset with samples queried from U, the length of Q can be 1 or a batch, and S is a superior that works to correctly label Q. As described above, active learning is an iterative process. At each iteration, the query set Q is added to T and removed from the unlabeled set U. It stops until the classified model is robust, or until the samples reach a fixed number.

### B. MAXIMUM MEAN DISCREPANCY (MMD)

Assume $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\} \in R^d$ are two datasets drawn randomly from a source dataset. Let p and q be two probability measures defined by X and Y respectively. The MMD is proposed to address the problem of whether the two probability measures p and q are similar. The principal potential of the MMD is to identify a function that assumes different expectations between different distributions so that when calculating the similarity between two distributions, it can empirically evaluate the quality of the MMD. Let F be a class of functions $g : \chi \to R$, and let p, q, x, y, X, Y be defined as above. According to [28], the MMD is:

$$MMD\left[\mathrm{G}, p, q\right] = \sup_{\mathrm{g}\in\mathrm{G}} \left(E_p\left[\mathrm{g}\left(x\right)\right] - E_q\left[\mathrm{g}\left(y\right)\right]\right) \quad (1)$$

Following [29], the empirical estimate of the MMD can be replaced by the empirical expectation computed for the samples X and Y. The empirical MMD can be defined as:

$$MMD\left[\mathrm{G}, X, Y\right] = \sup_{\mathrm{g}\in\mathrm{G}} \left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{g}\left(x_i\right) - \frac{1}{m}\sum_{i=1}^{m}\mathrm{g}\left(y_i\right)\right) \quad (2)$$

Clearly, when G is 'rich enough,' the MMD will disappear only if p = q. If G is 'restrictive' enough, the empirical estimate of the MMD will converge quickly to its expectation because the data size increases. It has been attested that the unit ball in the reproducing kernel Hilbert space (RKHS) with a characteristic kernel can satisfy both of the foregoing properties [30], [31]. Therefore, the unit ball in the RKHS can be used as the class of functions, then the MMD[G, X, Y] will be zero only if p = q. Let H be an RKHS. $\phi : X \to$H refers to the feature space mapping from the original to H, and F is a class of functions defined as the unit ball in RKHS. The MMD can be defined in RKHS, which can detect all the discrepancies between X and Y in RKHS. The empirical estimate of the MMD in RKHS is defined as

$$MMD_{\phi}\left[X, Y\right] = \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i\right) - \frac{1}{m}\sum_{i=1}^{m}\phi\left(y_i\right) \right\|_{\mathcal{H}}^{2} \quad (3)$$

## III. MULTI-CLASS ACTIVE LEARNING

Our approach is motivated by multi-label classification. Generally, active learning approaches follow the idea of uncertainty sampling, wherein samples on which the current classifier is uncertain are selected to be trained. The distance from the hyperplane of the classifiers has traditionally been used in research as representing uncertainty. However, it is difficult to extend this factor to multi-class classification due to the presence of multiple hyperplanes. Meanwhile, the design of an active learning approach is usually based on binary class; therefore, the label information is only 1 or −1. Therefore, it is difficult to balance the distance to the multiple hyperplanes. To overcome this challenge, in the proposed method, for an instance, a set of labels whose length is equal to the number of classes is given as the multi-label classification. However, it is different to multi-label classification, in which the set of labels of an instance may have several labels equal to 1. In our proposed IR-AL, the set of labels for an instance is simply that one of the labels is equal to 1. An example is shown in Fig.2. The label information will be sufficiently used.

Suppose we are given a dataset $\mathrm{D} = \{x_1, x_2, \ldots, x_n\} \in R^d$, and it is randomly split into two datasets: the unlabeled dataset $\mathrm{D}_u = \{x_1, x_2, \ldots, x_u\}$ and the labeled dataset $\mathrm{D}_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$. Each $x_i$ in $\mathrm{D}_u$ and $\mathrm{D}_l$ is a $d$ dimensional feature vector. $\mathrm{D}_u$ is the candidate dataset and $\mathrm{D}_l$ is initially empty or has limited labeled samples for active learning. The $y_i$ corresponding to $x_i$ in $\mathrm{D}_l$ is denoted as a set of labels which has been defined above, since we are focusing here on multi-class classification problems. Mathematically, $y_i = [y_{ik}]_{c\times 1} = [y_{i1}, y_{i2}, \ldots, y_{ic}]^T$, $c$ is the number of classes. If $x_i$ belongs to the class $c$, $y_{ic}$ is equal to 1, alternatively $y_{ic}$ is equal to −1. In our active learning method, we iteratively select one instance $x_s$ from the pool of unlabeled data $\mathrm{D}_u$ to query its label and to add the query instance to the labeled data $\mathrm{D}_l$. The goal is to gradually improve the model's accuracy. The symbols defined above will be used in the following discussion.

### A. UNCERTAINTY BY MINIMUM MARGIN

In order to mitigate the empirical risk for the active learning method based on the minimum margin, we must first review the margin-based active learning method itself. Let $f^*$ be the

| | | | | | |
|---|---|---|---|---|---|
| “airplane” | 1 | -1 | -1 | -1 | -1 |
| “butterfly” | -1 | 1 | -1 | -1 | -1 |
| “crab” | -1 | -1 | 1 | -1 | -1 |
| “crayfish” | -1 | -1 | -1 | 1 | -1 |
| “dalmatian” | -1 | -1 | -1 | -1 | 1 |

**FIGURE 2.** An example of the label definition of an instance for the proposed IR-AL.

classification model trained by the labeled samples:

$$f^* = \arg\min \sum_{i=1}^{l} l\left(Z_i, f\left(x_i\right)\right) + \lambda \left\|f\right\|_{\mathcal{H}}^2 \tag{4}$$

where H is RHKS endowed with kernel function $k\left(\cdot,\cdot\right) : R^d \times R^d \rightarrow R$, and $l\left(Z, f\left(x\right)\right)$ is the loss function $Z \in \{1, -1\}$. Given the classifier $f^*$, the margin-based approach chooses the unlabeled sample that is the most uncertain for the current classifier in $\mathrm{D}_u$:

$$x_s^* = \arg\min_{x \in \mathcal{D}_u} \left|f^*\left(x\right)\right| \tag{5}$$

Proposition 1 shows the equivalent form of the above formula in the active learning setting:

*Proposition 1:The uncertain instance in Eq.(5) can be rewritten as:*

$$x_s^* = \arg\max_{|Z_s|=1} \min_{f \in \mathcal{H}} \sum_{i=1}^{l} l\left(Z_i, f\left(x_i\right)\right) + l\left(Z_s, f\left(x_s\right)\right) + \lambda \left\|f\right\|_{\mathcal{H}}^2 \tag{6}$$

*Proof:* Define the object function by $\mathcal{U}\left(f\right)$:

$$\mathcal{U}\left(f\right) = \sum_{i=1}^{l} l\left(Z_i, f\left(x_i\right)\right) + \lambda \left\|f\right\|_{\mathcal{H}}^2$$

the process of proof is as follows:

$$\begin{aligned} x_s^* &= \arg\min_{x_s \in \mathcal{D}_u} \left|f^*\left(x_s\right)\right| \\ &= \arg\min_{x_s \in \mathcal{D}_u} \min_{f \in \mathcal{H}: \mathcal{U}(f) \le \mathcal{U}(f^*)} \left|f\left(x_s\right)\right| \\ &= \arg\min_{x_s \in \mathcal{D}_u, f \in \mathcal{H}: \mathcal{U}(f) \le \mathcal{U}(f^*)} \left|f\left(x_s\right)\right| \\ &= \arg\min_{x_s \in \mathcal{D}_u, f \in \mathcal{H}} \left|f\left(x_s\right)\right| + \Gamma \mathcal{U}\left(f\right) \\ &= \arg\max_{|Z_s|=1} \min_{x_s \in \mathcal{D}_u, f \in \mathcal{H}} l\left(Z_s - f\left(x_s\right)\right) + \Gamma \mathcal{U}\left(f\right) \end{aligned}$$

Let $\Gamma = 1$, the Eq.(5) is same to the Eq.(6), it is well to demonstrate that they are equivalent.

We extend the binary margin-based approach to the multi-class problem, first by considering the simple case of active learning that does not take into account the label correlation. For each label, we learn a classifier independently. The object function of the multi-class learning task with the least square can be represented as:

$$\begin{aligned} \max_{\left|\hat{y}_s\right|=1^c} \min_{f_k \in \mathcal{H}, x_s} & \sum_{i=1}^{l} \sum_{k=1}^{c} \left(y_{ik} - f_k\left(x_i\right)\right)^2 \\ & + \sum_{k=1}^{c} \left(\hat{y}_{sk} - f_k\left(x_s\right)\right)^2 + \lambda \sum_{k=1}^{c} \left\|f_k\right\|_{\mathcal{H}}^2 \end{aligned} \tag{7}$$

where $f_k$ is the classifier of the $k^{th}$ class, $y_{ik}$ is the true label of instance $x_i$ corresponding to the label $k$, $y_{sk}$ is the pseudo label of the query sample $x_s$ corresponding to label $k$, and $1^c$ is a vector of length $c$, with all entries 1. If we solve Eq.(7) with regard to $\hat{y}_s$ with fixed f and $x_s$, we minimize the worst-case risk introduced by the query sample. In this case, the pseudo-label $\hat{y}_{sk} = -sign\left(f_k\left(x_s\right)\right), k = 1, 2, \ldots, c$; therefore, the risk term becomes:

$$\begin{aligned} \min_{f_k \in \mathcal{H}, x_s} & \sum_{i=1}^{l} \sum_{k=1}^{c} \left(y_{ik} - f_k\left(x_i\right)\right)^2 \\ & + \sum_{k=1}^{c} \left(\hat{y}_{sk}^2 + 2\left|f_k\left(x_s\right)\right|\right) + \lambda \sum_{k=1}^{c} \left\|f_k\right\|_{\mathcal{H}}^2 \end{aligned} \tag{8}$$

which is still the upper bound of the true risk. For the sake of convenience, we use the linear model in the kernel space as the classifier, whose form is $f\left(x\right) = w^T \phi\left(x\right)$. $\phi\left(x\right)$ is the feature mapping function as described above. The informative information can be measured by the objective function:

$$\begin{aligned} \min_{w, x_s} & \sum_{i=1}^{l} \sum_{k=1}^{c} \left(y_{ik} - w_k^T \phi\left(x_i\right)\right)^2 \\ & + \sum_{k=1}^{c} \left(\left|w_k^T \phi\left(x_s\right)\right|^2 + 2\left|w_k^T \phi\left(x_s\right)\right|\right) + \lambda \sum_{k=1}^{c} \left\|w_k\right\|_{\mathcal{H}}^2 \end{aligned} \tag{9}$$

Let $w = [w_1, w_2, \ldots, w_c]$ be the coefficient matrix. The above formula can be rewritten as:

$$\min_{w,x_s} \sum_{i=1}^{l} \left\| y_i - w^T \phi(x_i) \right\|_2^2 + \left\| w^T \phi(x_s) \right\|_2^2 + 2 \left| w^T \phi(x_s) \right| + \lambda \|w\|_F^2 \quad (10)$$

Using a kernel form $w_k = \sum_{x_j \in T} \theta_{jk} \phi(x_j)$, and let $\theta_c = [\theta_{ik}]_{l\times 1} = [\theta_{1c}, \theta_{2c}, \ldots, \theta_{lc}]^T$, $\phi(T) = [\phi(x_1), \phi(x_2), \ldots, \phi(x_l)]^T$, $x_k \in T$; therefore, $w_k = \theta_k^T \phi(T)$, and the coefficient matrix can be reformed by $\theta$:

$$w = \left[ \theta_1^T \phi(T), \theta_2^T \phi(T), \ldots, \theta_c^T \phi(T) \right] = \theta^T \phi(T)$$

where $\theta = [\theta_1, \theta_2, \ldots, \theta_c]$ is the coefficient matrix of $w$ in the kernel space. Define $Y = [y_k]_{l\times c} = [y_1, y_2, \ldots, y_l]^T$ as the label matrix, $R \in R_+^{c\times c}$ is an incidence matrix between the labels. Meanwhile, a function $vec()$ is introduced to convert a matrix into a vector along the column. Hence, the objective function is modified by taking into account the label correlation

$$\begin{aligned} \min_{\theta} & \left\| vec(Y) - \left( vec(\theta) \left( R \otimes K_{\mathcal{D}_l \mathcal{D}_l} \right) \right)^T \right\|_2^2 \\ & + \left\| vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l s} \right) \right\|_2^2 + 2 \left| vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l s} \right) \right| \\ & + \lambda vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l \mathcal{D}_l} \right) vec(\theta) \end{aligned} \quad (11)$$

where $\otimes$ is the Kronecker product between the matrices, and $K$ is the kernel matrix with its elements $K_{ij} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.

### B. DIVERSITY (REPRESENTATIVENESS) BY MAXIMUM MEAN DISCREPANCY (MMD)

In the proposed framework, the MMD is adopted to measure the representative information that is the evaluation of the distribution difference between two sets of samples. Active learning can constrain the distribution of the labeled and query samples and make it as similar to the overall sample distribution as possible. According to the above description about the MMD, it can be empirically calculated in active learning as:

$$MMD_\phi(\mathcal{D}_l \cup x_s, \mathcal{D}_u / x_s) = \left\| \frac{1}{l+1} \sum_{x_i \in \mathcal{D}_l \cup x_s} \phi(x_i) - \frac{1}{u-1} \sum_{x_j \in \mathcal{D}_u / x_s} \phi(x_j) \right\| \quad (12)$$

Actually, $x_s$ is the target sample which it is the goal of the proposed active learning method to query. Therefore, the above formula should be defined as an alternative representation which can select $x_s$ from the unlabeled dataset $D_u$ by optimization. According to [12], the MMD can also be represented as:

$$\min_{\alpha:\alpha_i \in \{0,1\}, \alpha^T 1 = 1} \frac{1}{2} \alpha^T k_1 \alpha - \frac{l+1}{l+u} \mathbf{1}_l k_2 \alpha + \frac{u-1}{l+u} \mathbf{1}_u k_3 \alpha + const. \quad (13)$$

where $\alpha$ is an indicator vector of length $u$. Note that if $x_s$ is selected in the unlabeled dataset $D_u$, then the corresponding $\alpha_s$ will be 1; otherwise, $\alpha_i$ will be 0. 1 is a vector of the same to the dimension $\alpha$ with all entries 1. $1_l$ and $1_u$ are vectors of lengths $l$ and $u$ respectively, with all elements 1. The other terms are as follows: K is the kernel matrix with its element described above; $k_1 = k_3 = K_{\mathcal{D}_u \mathcal{D}_u}$; $k_2 = K_{\mathcal{D}_l \mathcal{D}_u}$; $K_{AB}$ denotes the kernel matrix between A and B. Since the second and third terms have the same variable, they can be simplified as a form of standard quadratic programming (QP):

$$\min_{\alpha:\alpha_i \in \{0,1\}, \alpha^T \mathbf{1} = 1} \frac{1}{2} \alpha^T K_1 \alpha + K_2 \alpha \quad (14)$$

where $K_1 = K_{\mathcal{D}_u \mathcal{D}_u}$, $K_2 = \frac{u-1}{l+u} \mathbf{1}_u k_3 - \frac{l+1}{l+u} \mathbf{1}_l k_2$.

### C. THE HYBRID INFORMATIVE AND REPRESENTATIVE MULTI-CLASS ACTIVE LEARNING METHOD

Therefore, a hybrid informative and representative information multi-class active learning method (IR-AL) is proposed. The framework of the proposed method can be represented as follows:

$$\begin{aligned} \min_{\substack{\theta \\ \alpha:\alpha_i \in \{0,1\}, \alpha^T 1 = 1}} & \left\| vec(Y) - \left( vec(\theta) \left( R \otimes K_{\mathcal{D}_l \mathcal{D}_l} \right) \right)^T \right\|_2^2 \\ & + \left\| vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l s} \right) \right\|_2^2 + 2 \left| vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l s} \right) \right| \\ & + \lambda vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l \mathcal{D}_l} \right) vec(\theta) + \beta \left( \frac{1}{2} \alpha^T K_1 \alpha + K_2 \alpha \right) \end{aligned} \quad (15)$$

where $\beta$ is the trade-off weight for balancing the informative and representative information. According to the discussion about measuring the informative and representative information, the empirical risk of the object function as described above approximates to the true risk upper bound under the original distribution. It is not difficult to imagine that the query sample $x_s$ is the bond between the two parts. Therefore, the $x_s$ in terms of informative information can also be selected with the indicator vector $\alpha$ from the unlabeled dataset $D_u$. Hence, the above formula can be reformulated as:

$$\begin{aligned} \min_{\substack{\theta \\ \alpha:\alpha_i \in \{0,1\}, \alpha^T 1 = 1}} & \left\| vec(Y) - \left( vec(\theta) \left( R \otimes K_{\mathcal{D}_l \mathcal{D}_l} \right) \right)^T \right\|_2^2 \\ & + \sum_{i=1}^{u} \alpha_i \left( \left\| vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l i} \right) \right\|_2^2 \right. \\ & \left. + 2 \left| vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l i} \right) \right| \right) \\ & + \lambda vec(\theta)^T \left( R \otimes K_{\mathcal{D}_l \mathcal{D}_l} \right) vec(\theta) + \beta \left( \frac{1}{2} \alpha^T K_1 \alpha + K_2 \alpha \right) \end{aligned} \quad (16)$$

Intuitively, the above problem is not convex. In this way, alternating optimization is adopted [32]. If $\alpha$ is fixed, the representative term is a constant and the above objective is a

problem of finding the best classifier with the labeled dataset $D_l$ and the query sample $x_s$.

$$\min_{\theta} \left\| vec(Y) - \left(vec(\theta)\left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right)\right)^T \right\|_2^2 + \left\| vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l s}\right) \right\|_2^2 + 2\left| vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l s}\right) \right| + \lambda vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right) vec(\theta) + const \quad (17)$$

The above formula can be solved by the alternating direction method of multipliers (ADMM) [33]. If $\theta$ is fixed, the terms in the informative information part will be constant, except the term that contains the indicator $\alpha_k$. The objective becomes:

$$\min_{\alpha:\alpha_i\in\{0,1\},\alpha^T 1=1} \times \sum_{i=1}^{u} \alpha_i \left( \left\| vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l i}\right) \right\|_2^2 + 2\left| vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l i}\right) \right| \right) + \beta\left(\frac{1}{2}\alpha^T K_1 \alpha + K_2 \alpha\right) + const \quad (18)$$

Here, we define:

$$K_3(i) = \left\| vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l i}\right) \right\|_2^2 + 2\left| vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l i}\right) \right|$$

The above objective function can be rewritten as:

$$\alpha^* = \min_{\alpha:\alpha_i\in\{0,1\},\alpha^T \mathbf{1}=1} \frac{\beta}{2}\alpha^T K_1 \alpha + (\beta K_2 + K_3)\alpha + const \quad (19)$$

This is a standard QP form of the indicator vector $\alpha$. If the constraint condition $\alpha : \alpha_i \in \{0, 1\}, \alpha \mathbf{1}^u = 1$ is restricted for the QP. The time cost will be expensive in terms of finding the best value for the QP. Therefore, we relax $\alpha_k$ from 0 to 1. The value in $\alpha^*$ which is closest to 1 will be set as 1; otherwise, it will be set as 0.

### D. THE SOLUTION

In this section, we will discuss in detail the process of solving our proposed method. According to [32], the aim of alternating optimization is to solve a non-convex problem by changing one kind of variable and fixing the others; this procedure is alternated for all variables until the convergence condition is satisfied. Hence, two steps are required to solve (16) by means of alternating optimization. As described above, first, if $\alpha$ is fixed, Eq.(17) can be solved by employing ADMM to calculate the best $\theta$. Second, $\theta$ is fixed. As in the first step, Eq.(18) is solved by employing QP to solve $\alpha$.

First, as the indicator vector $\alpha$ is fixed, objective (16) becomes (17). For the sake of computational simplicity, we define the incidence matrix R as an identity matrix and introduce an auxiliary variable $a = vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l s}\right)$. The objective function (17) becomes:

$$\min_{\theta} \left\| vec(Y) - \left(vec(\theta)\left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right)\right)^T \right\|_2^2 + \|a\|_2^2 + 2|a| + \lambda vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right) vec(\theta) + const$$

$$s.t.\ a - vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l s}\right) = \mathbf{0}, \quad \forall x_s \in \mathcal{D}_u \quad (20)$$

where 0 is a vector of length $c$, with all entries 0. Construct the augmented Lagrangian of Eq.(20) as:

$$L_\rho = \left\| vec(Y) - \left(vec(\theta)\left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right)\right)^T \right\|_2^2 + \|a\|_2^2 + 2|a| + \lambda vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right) vec(\theta) + \left(a - vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l s}\right)\right)\xi^T + \frac{\rho}{2}\left\| a - vec(\theta)^T \left(R \otimes K_{\mathcal{D}_l s}\right) \right\|_2^2$$

According to [33], the updating rules can be obtained as follows:

$$\theta^{k+1} = B^{-1} b^T$$

with

$$B = \left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right)^2 + \frac{\rho}{2}\left(R \otimes K_{\mathcal{D}_l s}\right)\left(R \otimes K_{\mathcal{D}_l s}^T\right) + \lambda\left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right),$$

and

$$b = vec(Y_L)\left(R \otimes K_{\mathcal{D}_l\mathcal{D}_l}\right) + \frac{1}{2}\xi^k\left(R \otimes K_{\mathcal{D}_l s}^T\right) + \frac{\rho}{2}a^k\left(R \otimes K_{\mathcal{D}_l s}^T\right)$$

$$a^{k+1} = sign(v)\left(|v| - \omega\right)_+$$

with

$$v = \frac{\rho\left(\theta^{k+1}\right)^T \left(R \otimes K_{\mathcal{D}_l s}\right) - \xi^k}{\rho + 2},$$

$$\omega = \frac{2}{\rho + 2}, \quad (q)_+ = \max(0, q)$$

$$\xi^{k+1} = \xi^k + \rho\left(a^{k+1} - \left(\theta^{k+1}\right)^T \left(R \otimes K_{\mathcal{D}_l s}\right)\right)$$

By employing AMDD to solve $\theta$ in the first step, $\theta$ is therefore foregone in the second step. The objective then becomes Eq.(19); there is a standard QP problem for $\alpha$. This problem can be solved using standard QP toolboxes such as MOSEK[1] or the function quadprog[2] in MATLAB. If the two steps all converge, for computing $\alpha$, the unlabeled sample in $D_u$ corresponding to the position of the largest element in $\alpha$ is the query sample $x_s$.

## IV. EXPERIMENTS

To investigate the proposed method's performance, we compare the proposed method with the following three state-of-the-art active learning methods in our experiments:

- QUIRE: a min-max-based active learning method [11], an approach which queries both informative and representative information instances.
- Adaptive: an active learning method which combines the uncertainty information and the representative information with product [19]. This is an approach for querying informative and representative instances.

**TABLE 1. Characteristics of the datasets, including the numbers of the corresponding features and samples.**

| Dataset | #Feature | #Instance |
|---|---|---|
| balance | 4 | 625 |
| iris | 4 | 150 |
| vehicle | 18 | 846 |
| wine | 13 | 178 |
| waveform | 21 | 5000 |
| vowel | 10 | 528 |
| australian | 14 | 690 |
| glass | 9 | 214 |
| ionosphere | 34 | 351 |
| image | 18 | 2086 |
| vote | 16 | 435 |
| landsat | 36 | 2000 |

**TABLE 2. Win/tie/loss counts of the proposed method versus the other state-of-the-art methods based on a paired t-test at a 95% significance level.**

| Dataset | Vs QUIRE | Vs Adaptive | Vs MP-AL |
|---|---|---|---|
| image | 15/5/5 | 25/0/0 | 25/0/0 |
| ionosphere | 17/8/0 | 25/0/0 | 24/1/0 |
| iris | 7/18/0 | 25/0/0 | 15/5/0 |
| landsat | 25/0/0 | 23/2/0 | 21/4/0 |
| vehicle | 25/0/0 | 25/0/0 | 24/1/0 |
| vote | 11/11/3 | 17/6/2 | 20/2/3 |
| vowel | 20/5/0 | 19/2/4 | 14/3/8 |
| waveform | 17/6/2 | 25/0/0 | 25/0/0 |
| wine | 5/17/3 | 3/22/0 | 20/5/0 |
| australian | 13/12/0 | 17/8/0 | 16/9/0 |
| balance | 13/9/3 | 23/2/0 | 24/1/0 |
| glass | 15/7/3 | 21/4/0 | 19/1/5 |

- MP-AL: an active learning method based on marginal probability distribution matching [12], an approach which prefers representative instances.

### A. SETTINGS

In our experiments, 12 datasets are used from UCI benchmarks. The characteristics of these datasets are summarized in Table 1. *vote*, *ionosphere*, *image,* and *australian* are benchmark datasets designed with binary class. Since many active learning methods are designed with binary class, in order to demonstrate the effectiveness of the proposed method, we conduct our experiments on these four benchmark binary class datasets. *balance, iris, vehicle, wine, waveform, vowel, glass,* and *landsat* are multi-class datasets with more than two classes. We randomly divide each dataset into two parts with percentages of 60% and 40%. The 40% is used as the test data and the other part is used as the unlabeled data for active learning. We assume that no labeled data is available at the very beginning of active learning. For the adaptive method, the initial labeled data is needed. The initial labeled data is randomly sampled from the 60% portion with samples from each class. For each class, there is only one sample labeled and added to the initial labeled set; therefore, the number of the labeled data is only sufficient to train an initial classifier. At each iteration, an instance is selected to solicit its label and the classification model is retrained. Meanwhile, the retrained classification model is evaluated according to its performance on the test data. Since active learning is an iterative process, we stop our experiments for each dataset when the learning accuracy does not increase for any method. However, for some datasets, the experiments stop much earlier due to the limited samples, e.g. *vowel, ionosphere*. For each dataset, we run the process ten times with different random partitions of the dataset. Furthermore, for the compared methods, the parameters we used follow their original papers. In all experiments, for enhancing fairness, the LIBSVM [34] is used to train a classifier in all methods. The classification accuracy curve of the SVM classifier after each query is used for evaluation metrics. An RBF kernel is used in all kernel calculations. For the SVM classifier, the main parameters are the penalty coefficient $C$ and the kernel bandwidth $g$. We set the two parameters with empirical values with 100 and *1/d*, where $d$ is the dimension of the original data. In our proposed method, there are also two parameters: the regularization weight $\lambda$ and the trade-off parameter $\beta$. We set the regularization weight as $\lambda = 0.1$. The trade-off parameter $\beta$ is chosen from a candidate set {1, 2, 10, 100, 1000} by cross-validation. In the experiments, a QP problem needs to be solved; therefore, we adopt the MOSEK toolbox in our experiments as the solver. For the adaptive and QUIRE methods, the inverse of the kernel matrix should be calculated; hence, if the dataset is too large, when running the two methods, the memory of the computer must be high. However, we also want to maintain the good performance of these methods; therefore, in our experiments, relatively small benchmark datasets have been selected.

### B. COMPARISON WITH STATE-OF-THE-ART METHODS

Fig.3 shows the classification accuracy of the proposed method and the varied numbers of the query samples for the methods to which it is compared. Table 2 shows the win/tie/loss counts of the IR-AL compared with the other state-of-the-art methods.

Intuitively, we consider that the proposed method IR-AL performs much better than the compared methods. First, we observe that the performance of QUIRE is better than the other two compared methods. It works well for half of the benchmark data sets, but poorly for the others. Since QUIRE is a min-max margin-based approach and requires unlabeled data for semi-supervised learning in order to boost the learning performance, hence, we attribute its poor performance for some datasets of the QUIRE to the fact that the unlabeled data structure does not satisfy the semi-supervised learning assumptions. The adaptive method performs the worst of all the methods. This method combines the uncertainty and representativeness with the product; however, they are calculated separately. Meanwhile, the representative information is measured by using the Gaussian Process. Therefore, it requires a large amount of unlabeled data to evaluate the distribution. If the unlabeled data is insufficient for correctly evaluating the distribution, performance will be bad.

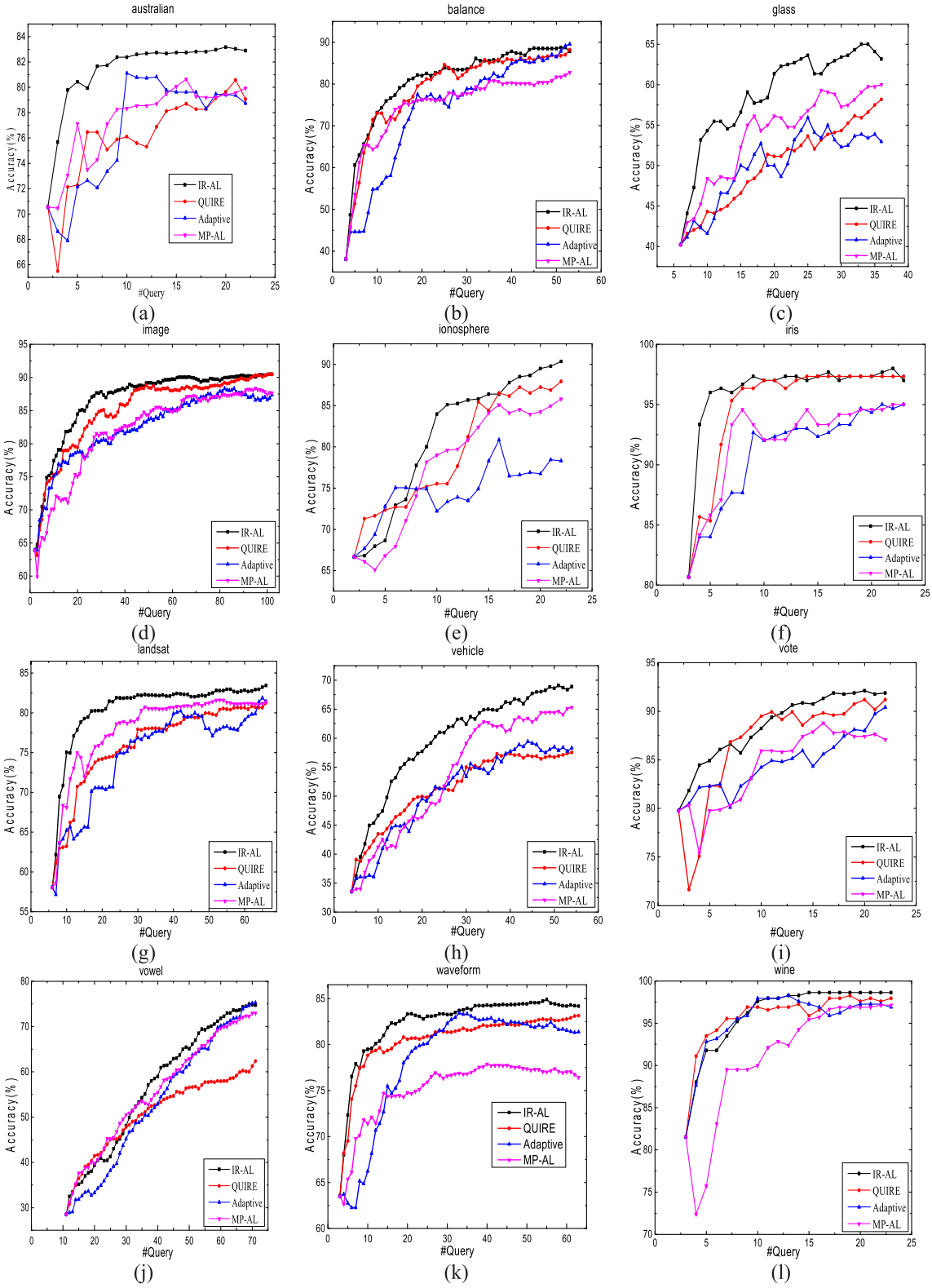



**FIGURE 3.** Comparison of different active learning methods for the 12 UCI benchmark datasets. The curve shows the learning accuracy over the queries. Each curve represents an average result of five runs.

Since the adaptive method needs to calculate an inverse matrix, the datasets used in this study are not large. This may be the most reasonable explanation why the adaptive method does not yield good performance for most of the datasets. The amount of unlabeled data is not sufficiently large. MP-AL performs better than the adaptive method; however, it only

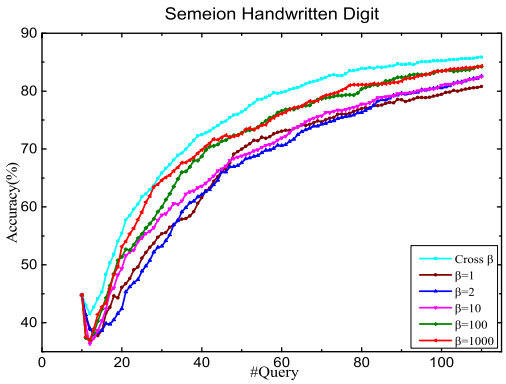



**FIGURE 4.** Performance comparison using different trade-off parameters on semeion handwritten digit data set for the proposed method. Each curve represents the average result of 10 runs.

uses the MMD as the query criterion. It also measures only the representativeness of the query instance, so it also needs a large amount of data. The reason why it performs better than the adaptive method may be that it has no requirement for unlabeled data distribution.

Finally, we observe that our method seldom performs worse than the compared methods. According to Fig.3 and Table 2, the proposed method of choosing informative and representative instances is successful for the multi-class datasets. These results demonstrate that both informative and representative information are significant in designing a good active learning approach. If a proper trade-off weight is used, it will boost the active learning performance.

### C. SENSITIVITY ANALYSIS

In our proposed method, the parameter $\beta$, which balances the informative and representative information in the formula, is a single factor which influences the experiment results [32]. In our experiments, we obtain the value of $\beta$ from a candidate set $\{1, 2, 10, 100, 1000\}$ in every run for our proposed method. To test the influence of this parameter, we show the results on a UCI benchmark dataset: a semeion handwritten digit which contains 1,593 instances with 256 features. The experimental settings are the same as the foregoing experiments except for the parameter $\beta$. We report our results in Fig.4. From the results, we can clearly observe that at the beginning of all the curves, there is poor performance; therefore, we consider that there is no immediate relationship with the parameter. Ignoring this phenomenon, we can observe that our method is more sensitive to the trade-off parameter $\beta$. When we use the candidate set by means of cross-validation in order to choose the best value for $\beta$, the curve is always higher than the curve with a fixed $\beta$. When $\beta$ has a small value such as $\{1, 2, 10\}$, there are no obvious changes for the three curves. When $\beta$ increases, performance also improves. However, at a certain interval, the curves do not change clearly. Although the changes are not clear at a specific interval, the larger values achieve greater performance. Therefore, the larger value for parameter $\beta$ is recommended. In this way, much more attention is paid to data distribution, which can boost active learning performance.

## V. CONCLUSION

In this study, we generalized the empirical risk principle to the active learning setting and proposed a novel multi-class active learning approach. By integrating uncertainty and diversity (representativeness), the proposed method can rapidly reduce the empirical risk as the data distribution is preserved. Meanwhile, we have made full use of the label information. Each instance has a set of labels, only one positive. In this situation, our proposed method will be more practical since most real-world problems are multi-class, and it also will boost active learning performance because it renders the samples being queried closest to all the classification hyperplanes of the binary class. The superior performance of our proposed method is demonstrated by the experiments conducted for 12 UCI benchmark datasets. Compared with other state-of-the-art methods, there is a limitation in our method. This is the trade-off parameter for balancing the uncertainty and diversity. Since the experimental results are sensitive to the parameter, this is a critical problem which restricts the applications of our method. Our future work aims to obtain the value of the parameter adaptively, which can make our method more practical. We also aim to develop our method to apply to multi-label classification and semi-supervised learning.

## REFERENCES

[1] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1648, 2009.
[2] O. M. Aodha, N. D. F. Campbell, J. Kautz, and G. J. Brostow, "Hierarchical subquery evaluation for active learning on a graph," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 564–571.
[3] N. Roy and A. McCallum, "Toward optimal active learning through Monte Carlo estimation of error reduction," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 441–448.
[4] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.
[5] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 150–157.
[6] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, no. 2, pp. 133–168, 1997.
[7] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Compression of hyperspectral remote sensing images by tensor approach," *Neurocomputing*, vol. 147, no. 1, pp. 358–363, Jan. 2015.
[8] B. Du and Y. Zhang, "Beyond the sparsity-based target detector: A hybrid sparsity and statistics based detector for hyperspectral images," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5345–5357, Nov. 2016.
[9] J. Wu, S. Pan, X. Zhu, C. Zhang, and S. Y. Philip, "Multiple structure-view learning for graph classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2703832.
[10] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 49–56.
[11] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.
[12] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Batch mode active sampling based on marginal probability distribution matching," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 741–749.

[13] J. Wu, S. Pan, X. Zhu, C. Zhang, and X. Wu, "Positive and unlabeled multi-graph learning," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 818–829, Apr. 2017.

[14] P. Donmez, J. G. Carbonell, and P. N. Bennett, "Dual strategy active learning," in *Proc. 18th Eur. Conf. Mach. Learn.*, 2007, pp. 116–127.

[15] C.-L. Li, C.-S. Ferng, and H.-T. Lin, "Active learning with hinted support vector machine," in *Proc. J. Mach. Learn. Res. Track*, vol. 25. 2012, pp. 221–235.

[16] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *Proc. 25th Eur. Conf. Inf. Retr. Res.*, 2003, pp. 393–407.

[17] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 79–86.

[18] B. Du *et al.*, "Exploring representativeness and informativeness for active learning," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 14–26, Jan. 2015.

[19] X. Li and Y. Guo, "Adaptive active learning for image classification," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 859–866.

[20] Z. Wang and J. Ye, "Querying discriminative and representative samples for batch mode active learning," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 158–166.

[21] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. Workshop Continuum Labeled Unlabeled Data Mach. Learn. Data Mining*, 2003, pp. 58–65.

[22] B. Du, M. Zhang, L. Zhang, R. Hu, and D. Tao, "PLTD: Patch-based low-rank tensor decomposition for hyperspectral images," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 67–79, Jan. 2017.

[23] J. Wu, S. Pan, X. Zhu, C. Zhang, and X. Wu, "Multi-instance learning with discriminative bag mapping," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2017.2788430.

[24] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–7.

[25] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[26] V. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[27] B. Du, Y. Zhang, and L. Zhang, "A hypothesis independent subpixel target detector for hyperspectral Images," *Signal Process.*, vol. 110, pp. 244–249, May 2015.

[28] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[29] A. Müller, "Integral probability metrics and their generating classes of functions," *Adv. Appl. Probab.*, vol. 29, no. 2, pp. 429–443, 1997.

[30] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Cambridge, MA, USA, 2008, pp. 489–496.

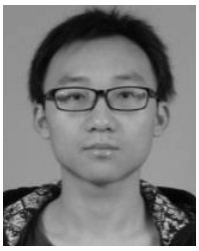
[31] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, Apr. 2010.

[32] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *Neural, Parallel Sci. Comput.*, vol. 11, no. 4, pp. 351–368, 2003.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

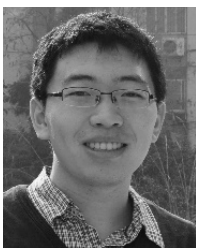
[35] B. Du and L. Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6844–6857, Nov. 2014.

**ZENGMAO WANG** received the B.S. degree in project of surveying and mapping from Central South University, Changsha, China, in 2013, and the M.S. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His current research interests include data mining and machine learning

**XI FANG** is currently pursuing the B.Sc. degree in information security with Wuhan University, Wuhan, China. He was a Senior Student with Wuhan University. In 2017, he has half a year research intern with the Department of Computer Science, Tsinghua University. He is currently a Research Member of the Research Group with the School of Computer Science. His research interests include computer vision, including object detection, image captioning, visual relationship detection, and active learning.

**XINYAO TANG** received the B.E. degree from the Computer Science and Technology Department, Ocean University of China, Qingdao, China, in 2015. She is currently pursuing the master's degree with the School of Computer Science, Wuhan University, Wuhan, China. Her research interests include pattern recognition.



**CHEN WU** (M'16) received the B.S. degree in surveying and mapping engineering from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2015. He is currently a Lecturer with the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include multitemporal remote sensing image change detection and analysis in multispectral and hyperspectral images.