IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2017, accepted October 5, 2017, date of publication November 8, 2017, date of current version February 1, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2761898

Active Learning for Visual Image Classification Method Based on Transfer Learning

JIHAI YANG[1,2], SHIJUN LI[1], AND WENNING XU[3]

[1]School of Computer Science, Wuhan University, Wuhan 430072, China
[2]Information & Telecommunication Branch, State Grid Jiangxi Electric Power CO, LTD, Nanchang 330077, China
[3]Institute of Mineral Resources, Chinese Academy of Geological Sciences, Beijing 100037, China

Corresponding author: Shijun Li (lishijunwhu@163.com)

ABSTRACT The active learning method involves searching for the most informative unmarked samples by query function, submitting them to the expert function for marking, then using the samples to train the classification model in order to improve the accuracy of the model and use the newly acquired knowledge to inquire into the next round, with the aim of getting the highest accuracy of classification using minimal training samples. This paper details the various principles of active learning and develops a method that combines active learning with transfer learning. Experimental results prove that the active learning method can cut back on samples redundancy and promote the accuracy of classifier convergence quickly in small samples. Combining active learning and transfer learning, while taking advantage of knowledge in related areas, could further improve the generalization ability of classification models.

INDEX TERMS Active learning, transfer learning, field adaptation, image classification.

I. INTRODUCTION

The arrival of the big data age has made it possible for technology services to improve through data mining and data analysis. However, it is fairly difficult to extract useful knowledge from the huge data, that's needed mining and find knowledge automatically. Therefore, Knowledge Discovery and Data Mining has been extensively studied. Traditional supervised learning methods are dependent on adequate marked samples to improve the generalization ability of classification models. However, in practical application, the number of marked samples is always limited, and the annotation of samples consumes a substantial amount of human resources, financial resources and time resources. For example, in the field of web page classification, search engines look for web pages using specific key words. At this time, this requires a judgement as to whether a particular page is related to the key words in question. However, surveys show that less than 0.0001% of web pages are annotated with subject labels. It is unrealistic to require that hundreds of millions of pages should be labeled [1]. This lack of marked samples has become a bottleneck for the development of artificial intelligence and other fields.

With further research into machine learning, data mining and other fields, and in the course of circumventing the lack of marked samples and other problems, researchers have put forward some effective results and solutions. Of these, active learning [2] is one of the most popular research fields. Unlike the traditional supervised learning method, active learning [4] obtains high quality knowledge selectively, selects the samples which contain the most information through the query function, and to give to expert to marked, use these sample training classifier models to improve model accuracy and use newly acquired knowledge to inform the next round of inquiry. In the course of the cycle, samples are actively selected and marked. The purpose of active learning is to use as few marked samples as possible to arrive at a high accuracy model while reducing the cost of marking up data. In machine learning problems where unmarked data is sufficient and lack of marked data or difficult to gain, active learning has good performance [3]–[10].

II. CURRENT SITUATION OF ACTIVE LEARNING

In the course of human learning, humans usually use existing experiences to learn new knowledge and rely on knowledge that has been previously obtained to sum up and accumulate experiences; in other words, experience and knowledge constantly interact. Similarly, machine learning simulates the human learning process, in that it uses existing knowledge to train the model to get new knowledge and thus arrives at a more accurate and useful new model through the constant accumulation of knowledge that corrects the model. Unlike traditional supervised learning, in which expert knowledge

is received passively, active learning can get high quality samples selectively. As such, active learning can reduce sample redundancy and achieve high classifier accuracy in small samples.

The active learning model can be defined as follows [2]:

$$A = (C, Q, S, L, U)$$

Here, C is a group or a classifier, L is a marked sample for model training, Q is the query function used for identifying the samples that have a large amount of information in an unmarked sample set U, and S is the supervisor, which can append the correct label to samples in U. The learner begins to study a small amount of initial marked samples L, selects a batch of the most useful samples through the query function Q, and asks the supervisor for a label, then uses the newly acquired knowledge to train the classifier and begin the next round of inquiries. Active learning is an iterative process that continues until it reaches a stop criterion. Its workflow is shown in Figure 1.

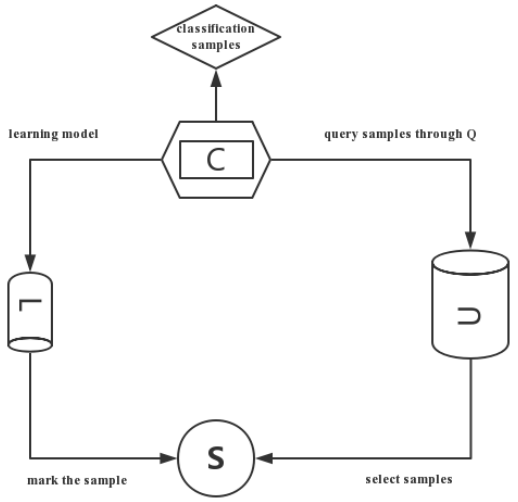


FIGURE 1. Active learning process.

Uncertainty-based Sampling (US) [11], [13], [32] is a simple and effective active learning method with a wide range of applications. In the active learning method, which is based on uncertainty, query function Q selects samples for which the classifier cannot determine the category. Because the least certain category samples often contain the most useful knowledge for the classifier, these samples are added to the training set to assist in the training of the classifier, which can improve the classifier's performance. However, in batch model active learning, which iterates each time to select a batch of samples, the query function only considers that uncertainty criteria may choose similar simples, causing information redundancy. One of the key questions in the batch model of active learning is how to select low-redundancy samples in order to provide as much information as possible for the classification model. Therefore, the query function must not only consider the uncertainty of samples, but also consider the difference between samples [6], [14], [15]. In recent years, moreover, transfer learning has been combined with active learning [17], [18], [34] in order to take advantage of knowledge from related fields while actively selecting target samples. This has also become an important research frontier in active learning.

III. UNCERTAINTY CRITERIA AND DIFFERENCE CRITERIA IN ACTIVE LEARNING

The query function Q of active learning is an important part of the active learning method, as it decides how to select samples and the quality of selected samples from the sample set. Therefore, the major differences in each active learning method center around the query function Q. There are two main types of query functions: uncertainty and difference. Uncertainty criteria mainly select samples for which the classifier cannot determine the samples of its category to provide as much information as possible for the classification model, while difference criteria select low-redundancy samples, also to provide the classification model with as much information as possible.

A. ACTIVE LEARNING METHOD BASED ON UNCERTAINTY CRITERIA

In each method of active learning, the use of uncertainty criteria is the most basic and intuitive strategy. Uncertainty criteria select samples for which the classifier cannot determine a category (since samples that cannot be readily allocated to a category always have the most useful knowledge for the classifier) and add these samples into the training set to assist in the training of the classifier, which can improve its performance.

1) MARGIN SAMPLING

Margin Sampling (MS) [19] is an active inquiry strategy based on *Support Vector Machines (SVM)*. In the SVM model, the main basis of determining the category to which the sample points belong is the distance of sample point to classification plane; therefore, the more sample points close to the classification plane, the more uncertainty exists as to which category belongs to. Based on this idea, MS select the sample points nearest to the classification plane to include in the training set, thus improving the accuracy of the classification model. In the case of multiclass classification, consider SVM that one against remains. If the total number of categories is n, then when training a second-class SVM between each category and the rest of the categories, n second-class SVMs will be trained in total. For each sample in an unmarked set U, select the distance from the nearest plane in n classification plane as a metric, then choose the minimum metric sample as the most uncertain sample. The query function is as follows:

$$x^{\mathrm{MS}} = \arg\min_{x_i \in U} \{\min_j |f_j(x_i, \alpha_j)|\} \quad (1)$$

Here, $f_j\left(x_i, \alpha_j\right)$ is the nth second-class SVM decision function value. As can be seen, for each sample in U, this strategy

first sets the distance of the sample to the nearest classification plane as the confidence. The smaller the confidence, the more easily the sample may be misclassified by mistake. Therefore, selecting the smallest sample of confidence to include into training set L could play a role in improving the performance of the classification model.

2) BASED ON MULTILEVEL SAMPLING OF UNCERTAINTY

The MS strategy, which considers the distance of samples to the nearest classification plane, could play a role in improving the classification model performance; however, because it only considers the sample to single classification plane distance, it may not be able to achieve a better effect in the more number of categories problem. Accordingly, to improve the MS basis, some academics have proposed a *Multiclass-Level Uncertainty (MCLU)* strategy [35]. Like MS, MCLU is a select strategy based on edge. In the case of multiclass classification, MCLU also considers the SVM that one against remains strategy. The difference is that MCLU considers the difference of sample points to the furthest two classification planes as a measure of the confidence of the sample. The one against remains strategy of SVM is ultimately based on the decision function value to determine the category of samples, so that the sample will be judged as the maximum decision function value of the class. Therefore, the smaller the difference of a sample to the furthest two classification planes, the lower the accuracy with which the one against remains strategy can judge what the class sample belongs to; that is to say, the greater the uncertainty of the sample. The query function is as follows:

$$\mathrm{x}^{MCLU} = \arg\min_{x_i \in U} \{c\,(x_i)\} \tag{2}$$

The $\mathrm{c}(x_i)$ in the above formula represents the confidence of sample x_i; that is, the difference of x_i and the furthest two classification planes. Its calculation process is as follows:

$$j = \arg\min_{j} f_j\,(x_i, \alpha_k) \tag{3}$$

$$j = \arg\min_{k \neq j} f_k\,(x_i, \alpha_k) \tag{4}$$

$$c\,(x_i) = f_j\left(x_i, \alpha_j\right) - f_k\,(x_i, \alpha_k) \tag{5}$$

Here, $f_j\left(\mathrm{x}, \alpha_j\right)$ is the decision function value of the j^{th} second SVM. It can be seen that the strategy first selects the largest two classification planes for this sample for each sample in unmarked set U, then calculates the difference of distance, chooses the minimum difference sample, and finally ensures selection of the most uncertain sample in each iteration.

3) Active Learning Method Combined with Uncertainty Criteria and Difference Criteria

In batch model active learning, the number of samples in each iteration round $b > 1$. In this case, simply considering the uncertainty of samples may cause extract to similar samples. These similar samples do not provide more information for model training and thus cannot improve the accuracy of the model effectively. Therefore, batch model active learning considers not only the uncertainty of the sample, but also the difference of samples, avoiding the extraction of multiple similar sample into the training set. In short, it considers the rules of uncertainty and difference together.

B. MARGIN SAMPLING-CLOSEST SUPPORT VECTOR

MS first measures the distance of sample to nearest classification plane as confidence, then chooses the minimum confidence sample to include into the training set, making it a sampling method based on uncertainty. One disadvantage of this method is that it is easy to cause data redundancy in batch model active learning. Therefore, some scholars have put forward an improved edge sampling method—*Margin Sampling-closest Support Vector (MS-cSV)* [21]. MS-cSV considers the support vector closest to candidate samples as a basis for candidate sample selection.

This method provides a series of support vectors in the SVM model: $\mathrm{SV} = (x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, that is, the training samples when the Lagrangian coefficient α is not zero. For each candidate sample vector q_i, we can choose the closest support vector (cSV) as follows:

$$cSV = \arg\min_{x_j \in SV} K\left(x_j, q_i\right) \tag{6}$$

Here, $\mathrm{K}\left(x_j, q_i\right)$ is the kernel function. If some candidate samples' cSV are the same, the sample closest to the classification hyperplane from them will be chosen. That is, the samples added in at each time cannot have the same cSV. Choice strategy is as follows:

$$q_b^{MS\_cSV} = \arg\min_{q_i \in U} \left(|f_j(q_i, \alpha_j)|\right) \cap cSV_b \neq cSV_l \tag{7}$$

where $l = 1, 2, \ldots, \mathrm{b} - 1$, is the subscript of the candidate samples that have been selected.

MS-cSV increases the limitation factor while selecting the sample closest to the classification hyperplane, meaning that the samples added in each iteration cannot have the same cSV. Obviously, in the case of selecting only one sample in each iterative round, the essence of MS and MS-cSV is the same, as they both choose the sample closest to the hyperplane. In batch model active learning, which makes multiple sample inquiries in each iterative round, MS-cSV could reduce redundant data as it considers the difference of samples, thus achieving better results.

1) MCLU-Enhanced Clustering-Based Diversity

In batch model active learning, combining uncertainty and difference could be an effective way to avoid data redundancy. Based on this idea, some scholars have proposed *MCLU-Enhanced Clustering-based Diversity (MCLU-ECBD)* [6]. This model adds differences in steps to avoid selection of high-similarity samples on an MCLU basis. The query function of MCLU-ECBD is based on the evaluation of uncertainty and difference in order to implement two coherent steps. If selected b samples are added into the training set in each iterative round, then the algorithm first uses MCLU

to select m(m>b) minimum degree of confidence samples in the uncertainty step, then uses kernel K-mean clustering to select the most different b samples from the m samples in the difference step. m/b means the balance of uncertainty and difference.

As MCLU is a method based on SVM, candidate samples are mapped to kernel space first, after which the confidence is evaluated. This time, to evaluate the difference between samples, the use of normal K-mean clustering will be insufficient to reflect the similarity of samples in kernel space; accordingly, K-mean clustering based on kernel space is used [22].

Kernel K-mean clustering iteratively divides m candidate samples into k=b classes (C1,C2, . . . ,Cb). In the first iteration, construct b classes (C1,C2, . . . ,Cb), and assign an initial class label to each sample. In the next iteration, each class selects a pseudo center ($\mu 1$, $\mu 2$, . . . , μb), so that the distance from each sample to each cluster center in kernel space can be calculated, after which each sample is assigned to the closest class. The Euclidean distance between $\emptyset(x_i)$ and $\emptyset(\mu_v)$, v=1, . . . ,b, is calculated as follows:

$$\begin{aligned} &D^2(\phi(x_i), \phi(\mu_v)) \\ &\quad = ||\phi(x_i) - \phi(\mu_v)||^2 \\ &\quad = ||\phi(x_i) - \frac{1}{|C_v|}\sum_{j=1}^{m}\delta(\phi(x_j), C_v)\phi(x_j)||^2 \\ &\quad = K(x_i, x_i) - \frac{2}{|C_v|}\sum_{j=1}^{m}\delta(\phi(x_j), C_v)K(x_i, x_j) \\ &\qquad + \frac{1}{|C_v|^2}\sum_{j=1}^{m}\sum_{l=1}^{m}\delta(\phi(x_j), C_v)\delta(\phi(x_l), C_v)K(x_j, x_l) \end{aligned} \quad (8)$$

$\delta\left(\emptyset(x_j), C_v\right)$ is an instruction function, j=1,2,. . .,m, v=1,2,. . .,b, only when x_j is assigned to C_v, $\delta\left(\emptyset(x_j), C_v\right) = 1$; otherwise, $\delta\left(\emptyset(x_j), C_v\right) = 0$, $|C_v|$ indicates that the total number of samples in class C_v, can be calculated as follows: $|C_v| = \sum_{j=1}^{m}\delta\left(\emptyset(x_j), C_v\right)$. $\emptyset(.)$ is a non-linear mapping function from primitive feature [16] space to high dimensional space, while K(..,..) is the kernel function. The kernel k mean algorithm can be summarized as follows:

Assign the initial value of $\delta\left(\emptyset(x_j), C_v\right)$, i=1,2,. .,m, v=1,2,. .,b, obtain b initial class{C1,C2,. .,Cb}.

Assign x_i to closest class:

$$\delta(\phi(x_i), C_v) = \begin{cases} 1, & D^2(\phi(x_i), \phi(\mu_v)) < D^2(\phi(x_i), \phi(\mu_j)), \\ & \forall j \neq v \\ 0, & other \end{cases} \quad (9)$$

From μ_v closest sample (calculated using formula (8)), the pseudo center of C_v was selected:

$$\eta_v = \arg\min_{x_i \in C_v} D^2\left(\phi(x_i), \phi(\mu_v)\right) \quad (10)$$

The algorithm has iterated to convergence, that is, no more samples are reassigned.

While kernel k-means algorithm get C1,C2, . . . ,Cb, selects the lowest degree of confidence samples from each class:

$$x_v^{MCLU-ECBD} = \arg\min_{\phi(x_i) \in C_v} \{c(x_i^{MCLU})\}, \quad v = 1, 2, \ldots, b \quad (11)$$

Finally, obtain b samples. These samples are labeled by the supervisor and added into the training set.

To summarize, MCLU-ECBD combines uncertainty and difference. At first, in the uncertainty steps, use MCLU to select samples which have the lowest degree of confidence, then consider the difference between samples, then k-mean clustering of samples which were selected in the first step in kernel space. We can think samples in each class are similarity, so to avoid data redundancy, select a minimum confidence of the sample to add into the training set from each class. Obviously, in cases where a sample is selected in each round of iterations, the second step of MCLU-ECBD does not work, so it is the same as the essence of MCLU. Moreover, in batch model active learning that selects multiple samples in each round of iterations, MCLU-ECBD can reduce redundancy effectively and improve classification accuracy.

IV. ACTIVE LEARNING COMBINED WITH TRANSFER LEARNING LUSION

Active learning training classification, through selecting the most information on the number of samples labeled to achieve the dual purpose of cost-savings and improvement of classification performance, effectively solves the problem of the insufficiently marked sample. However, in some fields of application, considering only the active learning method may cause a waste of marked data in related fields. For example, when categorizing blog documents, documents marked as 'news documents' will no doubt be of some help; ignoring news documents will cause a waste of data, thereby increasing the unnecessary burden of classification work. Consider the use of documents marked as news for the classification of blog documents as an instance of transfer learning. Transfer learning refers to migrating the accumulated knowledge from a related field to the target field. Unlike traditional supervised learning, in transfer learning, the training set and test set do not need to be subordinate to the same distribution. Therefore, in cases where there are adequate marked samples from a related field and a certain number of marked samples from the target field, combining transfer learning with active learning can maximize the use of marked samples from the field and reduce the work involved in marking samples from the target field, thus ultimately achieving the purpose of saving on costs and improving model performance.

1) Joint Optimization Framework for Transfer and Batch-model Active Learning

Active learning and transfer learning are two ways to solve the problem of insufficient marked samples: transfer learning solves this problem by acquiring knowledge from data

sources in a related field, while active learning is concerned with selecting as small as possible and a large amount of information samples to be manually annotated. In order to both use marked data in a related field and select large information samples in the target field, it has been proposed that transfer learning be combined with active learning to form a uniform framework—a *Joint Optimization Framework for Transfer and Batch-model Active Learning (JO-TAL)* [23]. The main characteristics of this framework are field adaptation on source domain data S, with active sampling on unmarked data in target domain U, to achieve the purpose of selecting training data which have similar probability distributions to the test data set. Suppose that for the source field and target sources, mark function or conditional probability P(y|x) are the same. As such, the above problem can be simplified as: domain adaptation on source field data set S to get field adaptation source data S_a while selecting subset Q from U to make the edge probability $P_{S_a \cup Q \cup L}(x)$ similar to $P_{U \setminus Q}$(x), while L represents marked data in the target field. This framework is described in detail below.

JO-TAL uses *MMD (Maximum Mean Discrepancy)* [25], [26], [36] to evaluate the difference in edge probabilities between two data sets. Supposing that source data or field adaptation source data S_a have n_s samples, target field unmarked data U have n_u samples, and target field marked data L have n_l samples, we hope that selecting a batch of query subset in U which includes b samples in each iteration will make the edge probability $P_{S_a \cup Q \cup L}(x)$ similar to $P_{U \setminus Q}$(x). MMD defines f between two sets of data as

$$f = \left\| \frac{1}{n_s + n_l + b} \sum_{j \in s_a \cup L \cup Q} \phi(x_j) - \frac{1}{n_u - b} \sum_{i \in U \setminus Q} \phi(x_i) \right\|_H^2 \tag{12}$$

Here, φ: X→H is a mapping from feature space X to high dimension space H. Selecting a subset Q of U to minimize the distribution difference between S_a∪L ∪ Q and U\Q is equivalent to selecting a Q to minimize f. We then defined a n-dimensional binary vector α, such that each component α_i indicates whether $x_i \in$ U is selected: if selected, $\alpha_i = 1$, otherwise, $\alpha_i = 0$. Domain adaptation achieved through re-weighting, a technology widely used in transfer learning [17], [27], [37], is repurposed to match the edge distribution of source field and target field data. In this respect, we defined another n_s-dimensional vector β, where each component β_i indicates weights of $x_i \in$ S. At this point, the problem is reduced to find the appropriate α and β to minimize the cost function:

$$\begin{aligned} \min \quad & \left\| \frac{1}{n_s + n_l + b} \sum_{i \in S} \beta_i \phi(x_i) + \sum_{j \in L} \phi(x_j) + \sum_{i \in U} \alpha_i \phi(x_i) \right. \\ & \left. - \frac{1}{n_u - b} \sum_{i \in U} (1 - \alpha_i) \phi(x_i) \right\|_H^2 \\ s.t.\ & \alpha_i \in \{0, 1\}, \quad \beta_i \in [0, 1],\ \alpha^T 1 = b \end{aligned} \tag{13}$$

Here, 1 is an all-1 vector of the same dimension as α. The first item of the above formula indicates the average of re-weighted source data, marked target data and the mapping feature of unmarked target data. For each sample x_i, if not selected, the corresponding component α_i is 0, so in the first item, the unselected sample x_i will not be counted as sum. The second item indicates the unmarked set U minus the mapping feature average of the selected query subset. The first limitation factor makes sure that each component of α has a value of 1 or 0, while the third limitation factor makes sure that only b samples are selected.

The formula (13) can be expressed as

$$\begin{aligned} \min \quad & \frac{1}{2} \alpha^T K_{u,u} \alpha + \frac{1}{2} \beta^T K_{s,s} \beta + \beta K_{s,u} \alpha - k_{u,u}^{\mathrm{T}} \alpha - k_{s,u}^{\mathrm{T}} \beta \\ & + k_{u,l}^{\mathrm{T}} \alpha + k_{s,l}^{\mathrm{T}} \beta + const \\ s.t.\ & \alpha_i \in \{0, 1\}, \quad \beta_i \in [0, 1],\ \alpha^T 1 = b \end{aligned} \tag{14}$$

Here, G is the kernel Gram matrix of $(n_s + n_u + n_l) * (n_s + n_u + n_l)$ on the source data S, unmarked data in target field U and marked data in target data L. While calculating, each data set is arranged in the order above. We can then use kernel K to define G, G (i, j) = K(x_i,x_j), and define c = $\frac{(n_s + n_u + n_l)}{n_u - b}$, such that

$$\begin{aligned} K_{s,s} &= \frac{1}{c^2} G(1 : n_s, 1 : n_s), \\ K_{u,u} &= G(n_s + 1 : n_s + n_u, n_s + 1 : n_s + n_u), \\ K_{s,u} &= \frac{1}{c} G(1 : n_s, n_s + 1 : n_s + n_u), \\ k_{u,u}(i) &= \frac{n_l + n_s + b}{c^2 (n_u - b)} \sum_{j=1}^{n_u} K_{u,u}(i, j), \\ k_{s,u}(i) &= \frac{n_l + n_s + b}{c^2 (n_u - b)} \sum_{j=1}^{n_u} K_{s,u}(i, j), \\ k_{s,l}(i) &= \frac{1}{c^2} \sum_{j=1}^{n_l} G(i, n_s + n_u + j), \\ k_{u,l} & \frac{1}{c} \sum_{j=1}^{n_l} G(i + n_s, n_s + n_u + j). \end{aligned}$$

Based on the above expression, we can observe the following: the first item of formula (14) ensures there is less similarity between the samples in selected query subset Q and avoids the data redundancy of query subset Q; the second item makes sure there is less similarity between the re-weighted source samples, avoiding the data redundancy of source data set S; the third item makes sure there is less similarity between the selected query samples and re-weighted source samples, avoiding information overlap; the fourth item makes the selected query samples and some unmarked samples similar, make sure the representative and typical; the fifth item makes the re-weighted source samples and some unmarked samples similar, ensuring the representativeness of source field data; the sixth item means less similarity between selected query samples and marked samples, ensuring the

difference of the selected data set; the seventh item means less similarity between reweighted source data and marked target data, ensuring the difference of the re-weighted source data set. Therefore, use of JO-TAL to choose samples can satisfy some of the requirements of both active learning and transfer learning, such as representativeness, difference, reduction of redundancy and information overlap.

Algorithm 1 JO-TAL

Input:
S: source field data;
L: marked data of target field;
U: unmarked data of target field;
b: number of samples in each iteration query;
β_{new} : source weight (input when the number of iterations is more than 1).
Output:
β_{new} : source weight (updated);
Q: target query set.
1: Calculate the H and f according to formula (15).
2: Solve formula (15) to calculate α and β.
3: Through the descending order of the components in α, obtain corresponding b samples, which is Q.
4: Update L, U: L $\leftarrow$ L $\cup$ Q, U $\leftarrow$ U\Q.
5: Update $\beta_{new}\beta_{new} \leftarrow \beta_{new} + \beta$: (when the number of iterations is more than 1).

The binary constraint for α_i makes formula (14) become an NP problem; therefore, we can make it become secondary planning through relaxing the limited factor, as follows:

$$X : X_i \in [0, 1], \quad X^T B = b^{0.5\sigma^T HX + f^T X}$$

where:

$$X = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}, \quad H = \begin{pmatrix} K_{s,s} & K_{s,u} \\ K_{s,u}^Y & K_{u,u} \end{pmatrix}, \; f = \begin{pmatrix} k_s, l - k_s, u \\ k_u, l - k_u, u \end{pmatrix}$$
$$B = \begin{pmatrix} o \\ I \end{pmatrix}, \quad I = 1_{n_u \times 1}, \; o = 0_{n_s \times 1}. \tag{15}$$

As this is a standard quadratic programming problem, there are many effective solutions. The key step of each iteration as algorithm-1 figure, β_{new} is source-weight vector, update in each iteration.

Active learning technology based on field adaptation: Query+ and Query−

Query+ and Query− [17] is a domain adaptation technology based on active learning. Its main principles are inquiring after a small amount of large information target domain samples in each iteration and deleting the samples not adapted to target domain distribution in the source field, combining the basic ideas of active learning and transfer learning. This method mainly uses the two query functions Query+ and Query− to achieve its aims: Query+ (Q+) selects the most information on the number of samples in the target field through the evaluation of uncertainty, while Query−(Q−) deletes the representative samples for target data in the source field from the current training set. Use the two query function, Q+ and Q− add new target samples into training while deleting the source samples in training set, so that the classification model is better adapted to target the field classification problem. In this way, the workload of manual marking is reduced at the same time.

The purpose of Q+ is to select a batch of largest information samples x^+ from the target field unmarked samples. These samples are then marked to be added into training set T (Here, training set T includes target field marked samples set L and source field data set S). For evaluation of uncertainty, use *Breaking Ties (BT)* [5]. BT is an active sampling strategy based on posterior probability and can be applied to selection in training set samples of Neural Networks, maximum likelihood classification and other probabilistic models. In probabilistic models, P(y|x) is the probability when the sample is x and label is y, that is, posterior probability. In general, the probabilistic classification model will be judged as the category with the highest probability of posterior probability, that is

$$\mathrm{i} = \arg\min_i P(y = \omega_i | x) \tag{16}$$

Therefore, for a sample x, the smaller the posterior probability difference between the two maximum different categories, the closer the probability of the two categories and the greater the uncertainty. This idea is fairly similar to MCLU, so the query function of BT algorithm is similar in form to the query function of MCLU:

$$x^{BT} = \arg\min_{x_i}(P(y = \omega_j | x_i) - P(y = \omega_k | x_i)) \tag{17}$$
$$j = \arg\min_j P(y = \omega_j | x_i) \tag{18}$$
$$k = \arg\min_{k \neq j} P(y = \omega_k | x_i) \tag{19}$$

Here, use maximum likelihood classifier [29] as classification mode. In the maximum likelihood classifier model, in order to avoid the number of samples less than or close to feature dimension cause to sample covariance matrix in the situation is strange or approximate strange, we can use regularization discriminant analysis [38] to estimate the covariance matrix.

The purpose of Q- is to delete x^-, which is the set of samples not adapted to target field distribution in the source field, and reduce the loss caused to classification model accuracy by this sample. Calculate the conditional probability of category using the Gaussian Probability Density Model [29], and consider the difference between using only the probability distribution calculated by source data S($T^{(0)}$) and the probability distribution calculated by current training set $T^{(i)}$. The query function is as follows:

$$x^- = \arg\min_{x \in T^{(0)}} \{p^{(0)}(x|\omega_l) - p^{(i)}(x|\omega_l)\} \tag{20}$$

Here, ω_l is the category to which x belongs. For each sample in source data set S, if the difference of conditional probability above is small, the probability distribution of

category ω_l calculated by source data S has not significantly changed; by contrast, if the difference above is great, the probability distribution of category ω_l from source field to target field has changed, and the corresponding x is no longer representative of target field distribution.

Combine Q+ and Q− together, add b^+ new target samples into the training set in each round of iteration, and delete the b^- source sample from training at the same time. We can go through the scale of source data set S, the relevance between source and target data and other factors to choose proportion b^-/b^+. If the source data and target data have a high relativity, we can select a smaller b^-; if relativity is low, select the higher b^-. The conclusion of Q+ and Q− is as follows:

Algorithm 2 Q+ and Q−

Input:
S: source field data;
L: marked data of target field;
U: unmarked data of target field.
b^+ : the number of target samples queried in each iteration;
b^- : the number of source samples deleted in each iteration.
Output:
Q^+ : target query set;
Q^- : source deleted set.
1: Use current training set $T^{(i)}(T^{(i)} = S \cup L)$ to train maximum likelihood classifier;
2: Go through Q^+ to select b^+ samples from U, which become Q^+;
3: Calculate the category conditional probability using $T^{(0)}(T^{(0)} = S)$ and $T^{(i)}$, respectively;
4: Go through Q- to select b^- samples in S, which become Q^-;
5: Update S, L, U : $S = SQ^-$, $L \leftarrow L \cup Q^+$, $U \leftarrow U \backslash Q^+$.

V. EXPERIMENTS

A. ACTIVE LEARNING METHOD EXPERIMENT

1) EXPERIMENTAL SETUP

In order to test the performance of the four active learning algorithms (MS, MCLU, MS-cSV and MCLU-ECBD), use the twelve UCI data sets widely used in the machine learning field as test data. The number of attributes, categories and samples of each data set are shown in Table 1.

This experiment uses random sampling comparison with the use of SVM training by all samples in the training set. The SVM model uses RBF kernel function, while the parameters use the optimal values obtained by cross-validation of the initial training set. Each data set is randomly divided into training set and test set by a ratio of 3:7. Each category randomly selects 5 samples as an initial training set, and each iteration selects 3 samples (b=3). In the uncertainty steps of MCLU-ECBD, the number of selected samples is 12 (m=4b). We performed 10 independent experiments in each data set, using a randomly selected initial training set every time, after which the results were averaged.

TABLE 1. 12 UCI data sets.

Data set	Attribute	Category	Sample
breast-cancer	9	2	263
diabetis	8	2	768
german	20	2	1000
heart	13	2	270
image	18	2	2086
iris	4	3	150
ringnorm	20	2	7400
splice	60	2	2991
thyroid	5	2	215
twonorm	20	2	7400
waveform	21	2	5000
wine	13	3	178

2) EXPERIMENTAL RESULTS AND ANALYSIS

We use paired t-test tests to evaluate the differences in the accuracy of the various methods. In this experiment, we compare the classification accuracy achieved by various methods in each independent experiment at a confidence level of 95% and record the instances of win/tie/loss. The results are shown in Table 2.

As we can see from Table 2, for each data set, active learning algorithm performance is better than that of the RS (random selection) algorithm. The advantage is particularly obvious in the performance of the *waveform* data set, but there is no significant difference between various active learning methods.

At the same time, we use the number of training set samples-classification accuracy curve to describe the trend of classification accuracy as the scale of training set increases. We can easily observe the performance differences between various algorithm through this curve. The experimental results for the 12 data sets are presented in Figure 2.

As we can see from Figure 2, on almost all data sets, the performances of active learning methods are better than those of the random sampling method. Especially in *breast-cancer, german, heart, image, iris, splice, thyroid* and*waveform,* the active learning method can greatly improve the classification accuracy compared to random sampling. However, because the number of categories in these data sets is low (a large number of data sets only include two types of samples), the superiority of MCLU relative to MS is not well reflected. This is because MS considers the distance from the candidate sample to the closest classification plane, while MCLU considers the difference between the two classification planes which are closest to the candidate sample. When the data set only includes two types of samples, there is essentially only one classification between the two types of samples,

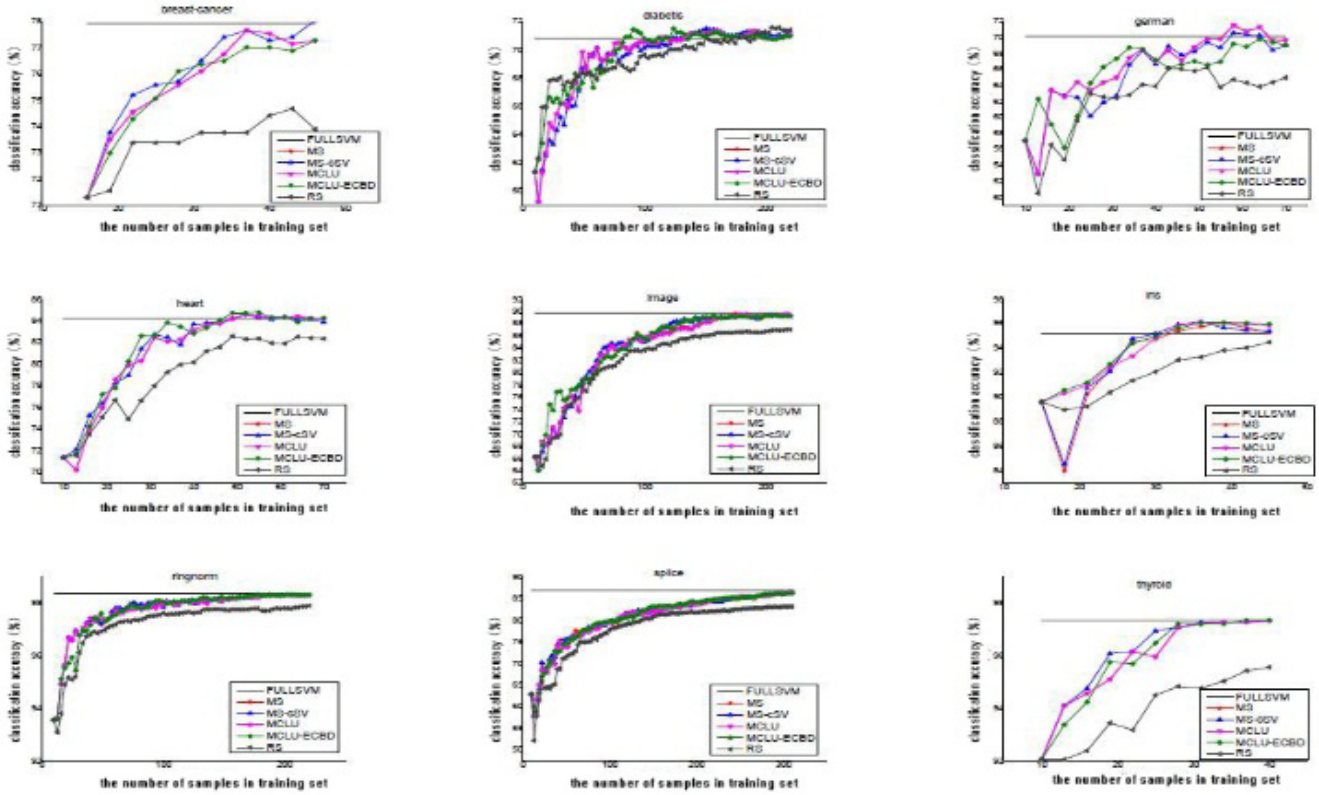


FIGURE 2. Experimental results of 12 UCI data sets.

so that in this case, there is not much difference between MS and MCLU. The algorithm that considered the difference of samples achieved relatively good performance on some data sets, such as in *breast-cancer* and *thyroid*, where MS-cSV achieved almost the best performance in each iteration. MCLU-ECBD achieved relatively good performance on *heart*. However, in general, for MS-cSV and MCLU-ECBD that considered the difference between samples compared to MS and MCLU, the advantage is not obvious, which could be due to the low redundancy between the samples in these data sets. In addition, as we can see from the early stages of the sample with less iterations, the active learning algorithm may be unstable, such as in *iris*. MS and MCLU-ECBD showed a significant decline of classification accuracy at the beginning of iteration. The classification accuracy of MCLU also declined in the first iteration.

The time complexity of various active learning methods in this section is presented in Table 3.

In order to verify the convergence speed of various active learning methods, experiments were carried out on *twonorm*, after which the calculation time required to achieve a certain accuracy was recorded. Results are presented in Table 4.

It can be seen that where the algorithm is relatively simple and the time complexity is low, the convergence speed of MS is fastest. However, this comparison method does not take into account the time it takes to mark the samples, as only the calculation time of inquired samples is considered; consequently, this method can to a certain extent reflect the convergence speed of various active learning methods, but cannot fully evaluate the performance of the algorithms.

B. ACTIVE LEARNING METHOD EXPERIMENT COMBINED WITH TRANSFER LEARNING

1) EXPERIMENTAL SETUP

Use 4 UCI data sets as test data to test the performance of the active learning method combined with transfer learning, as shown above. In order to ensure that the source data has a distribution difference with the target data, divide it according to a certain dimension for each data set. In *mushroom*, according to attribute *stalk-shape* to divide into two data sets, attribute stalk-shape for enlarging as source field data, attribute stalk-shape for tapering as target field data; in *enlarging-mushroom*, randomly select out 100 samples as source field sample. In *haberman*, according to attribute *Age of patient at time of operation* to divide, the value is between 45 and 65 as source data, and others as target field data [39]. In *kr-vs-kp*, according to attribute *Dwipd* to divide, *Dwipd* is source data for g and target data for 1. In *breast-cancer*, according to attribute *Clump Thickness* to divide, less than the average as source data, greater than average and the target data. The specific details are presented in Table 5.

TABLE 2. T-test table for the 12 UCI data sets.

Data set	MS VS MC LU	MS VS MS-cSV	MS VS RS	MCLU VS MCLU-ECBD	MC LU VS RS	MS-cSV VS RS	MCLU-ECBD VS RS
breast-cancer	14/7 2/14	12/6 5/23	58/3 3/9	16/71/13	58/3 3/9	64/2 6/10	58/32/10
diabetis	25/5 0/25	34/4 7/19	36/2 9/35	29/44/27	36/2 9/35	34/3 0/36	35/32/33
german	27/4 6/27	31/4 2/27	43/4 2/15	25/52/23	44/4 2/14	40/4 1/19	38/50/12
heart	19/6 2/19	19/6 1/20	53/4 2/5	14/61/25	53/4 2/5	56/3 8/6	58/42/0
image	35/3 4/31	35/3 0/35	64/2 1/15	32/30/38	65/1 9/16	66/1 8/16	66/23/11
iris	0/95/ 5	0/98/ 2	15/7 9/6	2/91/7	40/5 5/5	23/7 3/4	48/46/6
ringnorm	1/98/ 1	0/10 0/0	75/2 5/0	2/97/1	75/2 5/0	75/2 5/0	72/28/0
splice	18/6 5/17	18/6 0/22	99/1/ 0	15/54/31	98/2/ 0	97/3/ 0	94/6/0
thyroid	2/96/ 2	0/96/ 4	67/3 3/0	4/92/4	67/3 3/0	74/2 6/0	68/32/0
twonorm	11/8 7/2	11/8 0/9	38/5 9/3	10/86/4	30/6 5/5	38/5 7/5	27/65/8
wavefor m	16/6 0/24	17/6 6/17	100/ 0/0	19/56/25	100/ 0/0	100/ 0/0	100/0/0
wine	5/87/ 8	11/7 9/10	58/3 9/3	6/89/5	63/3 4/3	57/4 0/3	65/32/3

TABLE 3. Time complexity of various active learning methods.

Algorithm	Time complexity
MS	O(n)
MCLU	O(n)
MS-cSV	O(n^2)
MCLU-ECBD	O(n)

TABLE 4. Time required for various active learning methods to achieve a certain accuracy.

Accuracy (%) Time (s) Algorithm	94	95	96	97
MS	**0.008**	**0.023**	**0.044**	**0.649**
MCLU	0.029	0.108	0.194	1.549
MS-cSV	0.076	0.347	0.689	11.126
MCLU-ECBD	0.034	0.092	0.510	1.103

In addition to the two methods introduced by JO-TAL and Q+ and Q−, add the following algorithm as a comparison:

TABLE 5. Four UCI data sets.

Data set	Number of source data samples	Number of training data samples	Number of target test data samples
breast-cancer	121	20	122
haberman	136	60	110
kr-vs-kp	200	100	2105
mushroom	100	100	4508

1) *Joint Optimization based Transfer Learning and Rand Sampling. (JO-T-Rand)* [38]
   Operates mainly by solving the optimization problem in the target field for randomly selected samples.
2) *Optimization based Batch-mode Active Learning. (AL) [38]*
   This method only uses the data from the target field to train the classifier, reducing the difference between Q∪L and U\Q by solving optimization problems to select sample activity.
3) Q−
   This method makes a change to Q+ and Q−, randomly adding b^+ new target samples into the training set for each round, and at the same time, acting according to Formula (20) to delete b^- source samples from the training set.
4) BT
   Only uses the target data set to train the classification model, uses Formula (17) as a query function to inquire after samples in the target data set.

In the data sets *haberman*, *kr-vs-kp* and *mushroom*, select 10 samples as an initial training set; in data set *breast-cancer*, select four samples as an initial training set. In data sets *haberman* and *mushroom*, the number of target samples in each iteration b is 5, while in *kr-vs-kp* and *breast-cancer*, b is 3. Delete 10 samples for each round of iteration in all data sets. Using the SVM model, which uses the RBF kernel function, the parameters used are the optimal values obtained from the validation on the initial training set. Perform 10 independent experiments on each data set, using the random sampling initial training set every time, after which the results are averaged.

2) EXPERIMENTAL RESULTS AND ANALYSIS

We still use paired t-test tests to evaluate difference in the accuracy of the various methods. In this experiment, we compare the classification accuracy achieved by various methods in each independent experiment at a confidence level of 95% and record the instances of win/tie/loss. The results are shown in Table 6.

As we can see from Table 6, for most of the data sets, the performance of JO-TAL and Q+ and Q− where active learning and transfer learning is combined is better than that of the method JO-Rand and Q−, which only uses domain adaptation

TABLE 6. T-test table on four UCI data sets.

Data set	JO-TAL VS JO-Rand	JO-TAL VS AL	Q+ and Q- VS BT	Q+ and Q- VS Q-	JO-TAL VS Q+ and Q-
breast-cancer	5/94/1	64/36/0	0/100/0	1/97/2	15/85/0
haberman	60/38/2	91/9/0	100/0/0	0/100/0	0/0/100
kr-vs-kp	45/50/5	88/12/0	41/46/13	32/47/21	0/99/1
mushroom	13/87/0	17/83/0	10/90/0	10/90/0	0/64/36

and AL/BT (that only select initiatively). The experiment results proved that combined transfer learning and active learning can make the most of the data in related fields, reduce the work of target field mark samples, and finally achieve the goal of saving on costs and improving model performance. On *haberman* and *mushroom*, the performance of Q+ and Q− is better than JO-TAL, but on *breast-cancer*, the performance of JO-TAL is better than Q+ and Q−, while on *kr-vs-kp*, there is almost no difference between the performance of these two methods.

Similarly, in order to improve the observations of experimental results, we use the number of training sets samples-classification accuracy curve to describe the trend of classification accuracy as the scale of training set increases. The experimental results are shown in Figure 3.

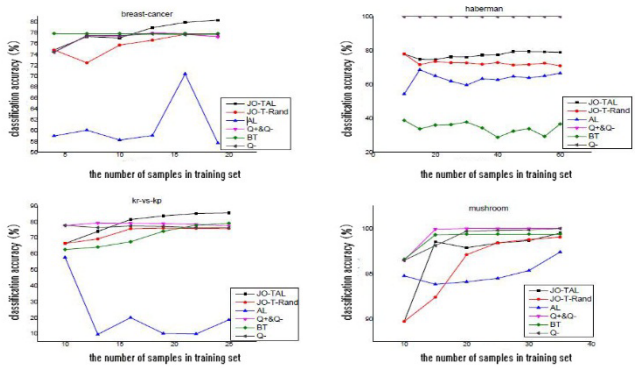


FIGURE 3. Experimental results for 12 UCI data sets.

As we can see from Figure 3, the method of JO-TAL and Q+ and Q− that combines active learning and transfer learning displays relatively good performance. On the data sets *breast-cancer* and *kr-vs-kp*, the performance of Q+ and Q− is better than JO-TAL at the beginning of the experiment; with an increase in the number of iterations, however, the classification accuracy of the Q+ and Q− model shows no obvious improvement, while the classification accuracy of JO-TAL gradually surpasses the classification accuracy of Q+ and Q−. This may be due to the parameters of Q+ and Q− model selection being inappropriate. On the data sets *haberman* and *mushroom*, the performance of Q+ and Q− is comparatively good; in the case of *haberman*, this may be because the similarity between source data and target data is fairly high, so that Q+ and Q− and Q− have maintained 100% classification from the beginning. On *mushroom*, Q+ and Q− quickly converge to 100%.

The time complexity of various algorithms in this section is presented in Table 7.

TABLE 7. Time complexity of various methods.

Algorithm	Time complexity
JO-TAL	O(n^2)
JO-T-Rand	O(n^2)
AL	O(n^2)
Q+ & Q-	O(n^2)
BT	O(n^2)
Q-	O(n^2)

In order to verify the convergence speed of the various active learning methods, an experiment was carried out on data set *mushroom* and the calculation time required to achieve a certain accuracy was recorded, as shown in Table 8.

TABLE 8. Time required by various methods to achieve a certain level of accuracy.

Time (s) / Accuracy(%) / Algorithm	96	97	98	99	100
JO-TAL	0.169	0.169	0.169	0.894	-
JO-T-Rank	0.060	0.060	**0.060**	-	-
AL	**0.057**	**0.057**	-	-	-
Q+ & Q-	0.517	0.517	0.517	0.517	**0.517**
BT	0.135	0.135	0.135	**0.135**	-
Q-	0.378	0.378	0.378	0.378	-

As we can see from Table 8, because the algorithm AL is relatively simple, the calculation speed is relatively fast while the accuracy requirements are low. However, due to there being no use of the samples of related fields, there was a failure to meet the high accuracy requirements. JO-T-Rand and BT can be seen to be faster at achieving higher accuracy requirements. As for Q+ and Q−, although the time required to complete a single iteration is long, it can meet higher accuracy requirements while the times of iteration are less. Similarly, it can to a certain extent reflect the convergence speed of various algorithms; however, because it does not take the time taken to mark samples into account, but only considers the calculation time of inquiring after samples, it cannot fully evaluate the performance of various algorithms.

VI. CONCLUDING REMARKS

This article first introduces various active inquiry strategies that employ uncertainty and difference. Uncertainty criteria use uncertainty as a measure of information about samples and select the sample with the highest degree of uncertainty to include in training in order to improve the performance of the model. Difference criteria consider the similarity between

samples and avoid the information redundancy caused by selected high-similarity samples. Combining the two methods of active learning and transfer learning solves the problem of insufficient marked samples, maximizes the use of marked data in related fields, reduces the work of marking samples in a related field, and ultimately achieves the goal of saving costs and improving the performance of the model. Finally, it is proven by experimental results that while the active learning method is faster than transfer learning, as it can converge quickly and reduce redundancy, the combination of active learning and transfer learning can take advantage of knowledge in related fields and improve the accuracy of the model.

For active learning, it is important to have effective stop criteria. If we can provide test data, then the iterative process of active learning can be achieved in cases where the classification accuracy reaches a steady state. However, as there may not be sufficient test data in the real world, we therefore hope to explore stop criteria that are not dependent on test data in the future. Moreover, as high-dimension input data may bring about ''dimension disaster'' during the inquiry process, we hope to explore some effective dimensionality reduction methods in the pretreatment stage to reduce the complexity of the query phase.

REFERENCES

[1] S. Sun and D. R. Hardoon, "Active learning with extremely sparse labeled examples," *Neurocomputing*, vol. 73, nos. 16–18, pp. 2980–2988, 2010.

[2] B. Settles, "Active learning literature survey," *Univ. Wisconsin Madison*, vol. 39, no. 2, pp. 127–131, 2009.

[3] L. Copa, D. Tuia, M. Volpi, and M. Kanevski, "Unbiased query-by-bagging active learning for VHR image classification," *Proc. SPIE, Image Signal Process. Remote Sens. XVI*, vol. 7830, Oct. 2010, doi: 10.1117/12.864861.

[4] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao, "Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1694–1707, Apr. 2017.

[5] T. Luo *et al.*, "Active learning to recognize multiple types of plankton," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3. 2004, pp. 478–481.

[6] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.

[7] E. Pasolli, F. Melgani, and Y. Bazi, "Support vector machine active learning through significance space construction," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 431–435, May 2011.

[8] B. Du, Y. Zhang, L. Zhang, and D. Tao, "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] A. Stumpf, N. Lachiche, J.-P. Malet, N. Kerle, and A. Puissant, "Active learning in the spatial domain for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2492–2507, May 2014.

[10] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6937–6956, Nov. 2014.

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

[12] 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.

[13] P. Ruiz, J. Mateos, G. Camps-Valls, R. Molina, and A. K. Katsaggelos, "Bayesian active remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2186–2196, Apr. 2014.

[14] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2014.

[15] B. Demir, L. Minello, and L. Bruzzone, "Definition of effective training sets for supervised classification of remote sensing images by a novel cost-sensitive active learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1272–1284, Feb. 2014.

[16] B. Du, X. Xiong, 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.

[17] C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4468–4483, Nov. 2012.

[18] C. Persello, "Interactive domain adaptation for the classification of remote sensing images using active learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 736–740, Jul. 2013.

[19] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 7th Int. Conf. Mach. Learn.*, 2000, pp. 839–846.

[20] B. Du *et al.*, "Exploring representativeness and informativeness for active learning," *IEEE Trans. Cybern.*, vol. 41, no. 1, pp. 14–26, Jan. 2017.

[21] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[22] K. M. Borgwardt *et al.*, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[23] R. Chattopadhyay *et al.*, "Joint transfer and batch-mode active learning," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 253–261.

[24] 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.

[25] J. Huang *et al.*, "Correcting sample selection bias by unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 601–608.

[26] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Planning Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[27] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning under covariate shift," *J. Mach. Learn. Res.*, vol. 10, pp. 2137–2155, Sep. 2009.

[28] B. Du and L. Zhang, "Target detection based on a dynamic subspace," *Pattern Recognit.*, vol. 47, no. 1, pp. 344–358, 2014.

[29] G. M. Foody, N. A. Campbell, N. M. Trodd, and T. F. Wood, "Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification," *Photogramm. Eng. Remote Sens.*, vol. 58, no. 9, pp. 1335–1341, 1992.

[30] B. Du and L. Zhang, "Random-selection-based anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1578–1589, May 2011.

[31] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," in *Proc. ICML*, 2000, pp. 111–118.

[32] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM active learning approach for image classification using spatial information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2217–2233, Apr. 2014.

[33] Z. Cui, X. Chen, J. Wu, V. S. Sheng, and Y. Shi, "Maximum classification optimization-based active learning for image classification," in *Proc. 7th Int. Congr. IEEE Image Signal Process. (CISP)*, Oct. 2014, pp. 759–764.

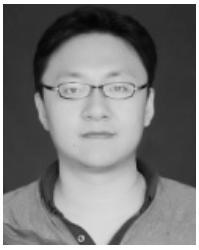
[34] X. Shi, W. Fan, and J. Ren, "Actively transfer domain knowledge," in *Machine Learning and Knowledge Discovery in Databases*. Antwerp, Belgium: Springer, 2008, pp. 342–357.

[35] A. Vlachos, "A stopping criterion for active learning," *Comput. Speech Lang.*, vol. 22, no. 3, pp. 295–312, 2008.

[36] B. K. Sriperumbudur *et al.*, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, Apr. 2010.

[37] J. D. Paola and R. A. Schowengerdt, "A detailed comparison of back-propagation neural network and maximum-likelihood classifiers for urban land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 4, pp. 981–996, Apr. 1995.

[38] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[39] T. Wang, B. Du, and L. Zhang, "A background self-learning framework for unstructured target detectors," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1577–1581, Nov. 2013.



JIHAI YANG received the M.S. degree in software engineering from Wuhan University, Wuhan, China, in 2009, and the B.S. degree in electronics and information engineering from Nanchang University, Nanchang, China, in 2004.

He is currently pursuing the Ph.D. degree with the School of Computer, Wuhan University. He is also an Engineer with the State Grid Corporation of China. He holds 18 patents. His major research interests include data mining, machine learning, pattern recognition, and information technology and telecommunications for the electric power system. He has received 11 provincial and ministerial level awards from China. He is currently a a member of specialized committee of the Chinese Society for Electrical Engineering and the telecommunication standard workgroup in China.

SHIJUN LI received the B.S. degree in mathematics from Peking Normal University, Beijing, China, in 1987, and the M.S. degree in computer science and the Ph.D. degree from Wuhan University, Wuhan, China, in 1998 and 2001, respectively.

He is currently a Professor with the School of Computer, Wuhan University, Wuhan. He has authored over 30 research papers, among them including International Conference on Web-Age Information Management (WAIM), International Conference on Web Information System Engineering, The Joint International Conferences on Asia-Pacific Web Conference and Web-Age Information Management, and the ACM/IEEE Joint Conference on Digital Libraries. His major research interests include data mining, machine learning and pattern recognition. He has served as a member of the program/organizing committees of three international conferences, as the Program Committee Chair for one international conference.

WENNING XU received the Ph.D. degree in agricultural informatization from the College of information and Electrical Engineering, China Agricultural University, Beijing, China, in 2011.

She is currently an Engineer with the Institute of Mineral Resources, Chinese Academy of Geological Sciences.