Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Interactive localized content based image retrieval with multiple-instance active learning

Dan Zhang[a,*], Fei Wang[a], Zhenwei Shi[b], Changshui Zhang[a]

[a]State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, P.R. China
[b]Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, P.R. China

ARTICLE INFO

Article history:
Received 31 May 2008
Received in revised form 4 January 2009
Accepted 2 March 2009

Keywords:
Multiple-instance active learning
Fisher information
Simple margin
Interactive localized content based image retrieval

ABSTRACT

In this paper, we propose two general *multiple-instance active learning* (*MIAL*) methods, *multiple-instance active learning with a simple margin strategy* (*S-MIAL*) and *multiple-instance active learning with fisher information* (*F-MIAL*), and apply them to the active learning in *localized content based image retrieval* (*LCBIR*). S-MIAL considers the most ambiguous picture as the most valuable one, while F-MIAL utilizes the fisher information and analyzes the value of the unlabeled pictures by assigning different labels to them. In experiments, we will show their superior performances in LCBIR tasks.

1. Introduction

Multiple-instance learning (*MIL*) can be viewed as a variation of the learning methods for problems with incomplete knowledge on the labels of the training examples (or *instances*). In the traditional *single-instance* learning, each training *instance* is assigned a discrete or real-valued label, while in the MIL settings, training patterns are given as *bags*, and each bag consists of some instances. The labels are assigned to bags, rather than instances. In a binary classification problem, a typical assumption for MIL is that a bag should be labeled as positive if at least one of its instances is positive; and negative if all of its instances are negative (there also exist some other MIL assumptions [2,5,8]). The goal of MIL is to learn a classifier that can predict labels of new bags or instances.

As another research branch of machine learning, *active learning* [1] is designed to reduce the labeling cost by the active interactions with users. In active learning, for each query, the learning machine selects a sample that is considered to be most valuable for labeling. Then the user will be asked to label the selected sample, and the classifier is retrained afterwards. Such a process will be repeated until a stopping criterion is satisfied. In this way, the performance of the classifier is expected to be high with only a few queries. If the queried object is a bag, rather than an instance, then it will turn to a *multiple-instance active learning* (*MIAL*) problem. In fact, this problem is very common in *localized content-based image retrieval* (*LCBIR*).

In LCBIR [2], what a user wants to retrieve is a region of interest (e.g. an object). In order to tell the retrieval system what he/she really wants, the user needs to provide several pictures. Each picture with the desired object on it is treated as a positive bag, while the other pictures are considered as negative ones. Then, after using image segmentation techniques to divide each picture into small patches, each patch in this picture represents an instance. However, the pictures provided by the user are always rare and we cannot get a satisfactory classifier by merely utilizing these pictures. To solve this problem, some researchers have considered using active learning to interact with the users and getting a satisfactory retrieval result with a few query times [3].

But almost all the active learning methods developed so far only consider the single-instance case. Therefore, in order to treat the active learning in LCBIR, we need to devise MIAL methods. So far, the MIAL problem has already arose some notices. In [4], the authors devised a MIAL algorithm based on MILES [5] and the simple margin strategy [31]. In [6], the authors also considered the MIAL problem, but with a different problem setting, where they intend to query the most valuable unlabeled instance in positive bags, rather than the most valuable unlabeled bag. It is true that their method is effective when we are allowed to query unlabeled instances. But for LCBIR, querying pictures (bags) would be more convenient than querying patches (instances).

To solve MIAL, we propose two general MIAL methods—*multiple-instance active learning with a simple margin strategy* (*S-MIAL*) and *multiple-instance learning with fisher information* (*F-MIAL*) to help choose the most valuable bag for query. Both S-MIAL and F-MIAL can be applied to the cases, when different MIL algorithms are employed

[*] Corresponding author. Tel.: +86 10 62796872; fax: +86 10 62786911.
E-mail address: dan-zhang05@mails.tsinghua.edu.cn (D. Zhang).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.03.002

to train the base classifiers. S-MIAL is also based on the *simple margin strategy*, with the method in [4] as its special case when the base learner is chosen to be MILES. It usually achieves a better performance than the random selection strategy. However, S-MIAL does not take into account that the benefits brought by treating the positive and negative bags separately. In other words, if the user gives a negative label to an unlabeled bag, then all the instances in this bag can be deemed as negative. Nevertheless, if this bag is labeled positive, we cannot simply assign this positive label to all the instances in this bag. We call this the asymmetric information carried by positive and negative bags. By utilizing the fisher information, we can handle this distinction and put forward F-MIAL.

2. Related works

In this section, we will briefly review some works that are closely related to this paper.

2.1. Multiple-instance learning

The notion of MIL was first introduced by Dietterich et al. [7] to deal with the drug activity prediction. After that, a lot of researchers have studied this problem. These algorithms can be roughly divided into three groups: the group that is specifically designed to solve multiple-instance problems, the group that tries to modify single-instance learning for MIL by introducing MIL constraints and the group that try to convert MIL to a single-instance problem and solve it using single-instance methods.

For the first group, the first MIL method is APR [7], which represents the target concept by an axis-parallel rectangle in the feature space. Maron and Lozano-Pérez [8,9] proposed a method called DD, which intends to find a concept point in the feature space that resembles positive instance most, and classify instances according to the distance between the instances and this concept point. Expectation–maximization (EM)-DD [2,10] combines EM with the DD formulation by using EM to search for the most likely concept.

As for the second group. Andrews et al. [12] proposed two methods based on SVM, one (mi-SVM) for the instance-based classification and the other (MI-SVM) for bag-level classification. Gehler and Chapelle applied deterministic annealing to the SVM formulation [13]. Gärtner et al. [14] proposed a kernel function directly for bags. Kwok and Cheung [15] extended their work by proposing a marginalized MI kernel to convert the MIL problem from an incomplete data problem to a complete data problem. Zhou et al. consider the multiple-instance multiple-label problem [16]. In their later work [17], they treat the MIL as a semi-supervised problem and try to maximize the margins for instances in the positive bags. In [18], the authors make some modification on the constraints of single-instance SVM and focus more on the positive bags with smaller sizes.

For the third group, DD-SVM [19] selects a set of prototypes from local maxima of DD function and then a SVM is trained based on the bag features summarized by these selected prototypes. In [5], bags are embedded into a feature space defined by instances, and a 1-norm SVM is applied to build the bag level classifiers.

These methods are quite effective for the MIL. However, they cannot actively interact with users to help improve their performance. Therefore, we try to exploit them in an interactive context and suggest an optimal way to use them iteratively.

2.2. Active learning

Active learning is a powerful machine learning approach that is designed to get satisfactory learning results with the minimum amount of labeled samples. For this purpose, many algorithms have been proposed to query the most valuable samples. These algorithms can be mainly grouped into two categories.

Algorithms in the first category tend to select a sample that is the closest to the current decision boundary for each query, which is also the most dubious sample for this classifier [20–23]. The theoretical basis of these algorithms are mainly on the maximum shrinking of the version space [24]. The motivation of the other kind of algorithms is to choose the most representative sample, i.e. query the unlabeled samples that can best represent the unlabeled data [25,26,1,27,28]. These algorithms are mainly based on the fisher information of the data distribution [1]. Recently, some combination algorithms have also been developed. The basic thought of these algorithms is to automatically choose different criterions for different cases [29,30]. Note that, in the context, we will not make any distinctions between the meanings of “sample” and “instance”, where the former is frequently used in active learning and the latter in MIL.

In this paper, all the above methods are designed for single-instance. However, as we have mentioned in the Introduction, many real-world application should be treated as multiple-instance problems. Hence, this also necessitates the need to devise MIAL methods and exploit active learning in an interactive context, with applications like image retrieval.

3. Problem statement and notations

Suppose we are initially given a set of N bags, $\{\mathbf{B}_i, i=1,2,\ldots,N\}$. The instances in the bag $\mathbf{B}_i$ are denoted as $\{\mathbf{B}_{i1},\mathbf{B}_{i2},\ldots,\mathbf{B}_{in_i}\}$, where n_i is the total number of instances in this bag. $y_i \in \{-1,+1\}$ is the class label for $\mathbf{B}_i$, and y_{ij} is the label of the instance $\mathbf{B}_{ij}$. In the typical MIL assumption, if the label of the ith bag, i.e., y_i, is $+1$, then at least one element of the label set $\{y_{i1},y_{i2},\ldots,y_{in_i}\}$ is $+1$. Otherwise, if the label of the ith bag, i.e., y_i, is -1, all the elements in the label set $\{y_{i1},y_{i2},\ldots,y_{in_i}\}$ are -1.

Among these bags, $l(0)$ bags are initially labeled, and denoted as $\{(\mathbf{B}_1,y_1),(\mathbf{B}_2,y_2),\ldots,(\mathbf{B}_{l(0)},y_{l(0)})\}$ while the other bags are unlabeled, and are referred to as $\{\mathbf{B}_{l(0)+1},\mathbf{B}_{l(0)+2},\ldots,\mathbf{B}_N\}$. Since the number of labeled bags, i.e., $l(0)$, is always small. Simply using $\{(\mathbf{B}_1,y_1),(\mathbf{B}_2,y_2),\ldots,(\mathbf{B}_{l(0)},y_{l(0)})\}$ to design the classifier cannot get a satisfactory performance. So, we consider using an interactive scheme to solve this problem. For the t-th query, an unlabeled bag $\mathbf{B}_{c(t)}$ with the largest criterion function value $Crit(\mathbf{B}_{c(t)})$ will be queried by the learning machine. After $y_{c(t)}$ is given by the user, $(\mathbf{B}_{c(t)},y_{c(t)})$ will be added to the labeled set. Then, the labeled set becomes $\{(\mathbf{B}_1,y_1),(\mathbf{B}_2,y_2),\ldots,(\mathbf{B}_{l(0)},y_{l(0)}),(\mathbf{B}_{c(1)},y_{c(1)}),\ldots,(\mathbf{B}_{c(t)},y_{c(t)})\}$, and the unlabeled set turns to: $\{\mathbf{B}_{l(0)+1},\mathbf{B}_{l(0)+2},\ldots,\mathbf{B}_N\}\backslash\{(\mathbf{B}_{c(1)}),\ldots,(\mathbf{B}_{c(t)})\}$ (“\” means ruling out). After that, the base classifier will be retrained based on the newly expanded labeled bags. The key problem for the above framework is how to design a proper criterion function $Crit(\mathbf{B}_i)$ so that the query times required to achieve a satisfied accuracy can be minimized. The main challenge for the design of the criterion function is the *label ambiguity* existed in multiple-instance setting. In the following sections, we will propose two general methods to devise this criterion function.

4. The proposed methods

In this section, our main focus is to design a proper criterion function $Crit(\mathbf{B}_i)$ for the t-th query. We consider two methodologies, i.e., S-MIAL and F-MIAL, to achieve this goal. S-MIAL is based on the *simple margin strategy*, which treats the bag closet to the current classification plane as the most valuable one. Compared with S-MIAL, F-MIAL is more complicated. It utilizes the fisher information, a value that indicates the uncertainty of the classifier, to measure the value

Table 1
S-MIAL.

```
Input:
1.  labeled bags: ${(\mathbf{B_1}, y_1), \ldots, (\mathbf{B_{l(0)}}, y_{l(0)})}$, where $l(0)$ is the number of the initially
    labeled bags,
2.  unlabeled bags: ${\mathbf{B}_{l(0)+1}, \ldots, \mathbf{B}_N}$, where $N$ is the number of bags,
3.  parameters for mi-SVM.
    For $t$:1 to TotalQueryTimes
      ForEach Bag $\mathbf{B}_i$ in the unlabeled set
        Calculate $Crit(\mathbf{B}_i)$ by Eq. (1).
      End ForEach
    Ask query $\mathbf{B_{c(t)}}$ for which $Crit(\mathbf{B}_{c(t)})$ is the largest and get the label $y_{c(t)}$.
    Add $(\mathbf{B}_{c(t)}, y_{c(t)})$ to the labeled set, and remove $\mathbf{B}_{c(t)}$ from the unlabeled set.
    Update Classifier $\mathscr{C}$ an SVM based multiple instance algorithm.
    End For
Output: The final classifier $\mathscr{C}$
```

of the unlabeled bags. We will elaborate these two methods in this section.

4.1. MIAL with a simple margin strategy (S-MIAL)

In [31], Tong demonstrated that, in traditional single-instance learning, when SVM [32] is used as the base learner, we should query an example that can bisect the version space (The version space is in fact a space for parameters of the classifier. Each example in the feature space can be represented by a hyperplane that cuts through the version space.) Based on this theory, they put forward *SVMactive* with a simple margin strategy, which considers the instance $\mathbf{x}_i$, whose output $f(\mathbf{x}_i)$ is the closest to 0, as the most valuable one.

A natural extension of this method to the multi-instance scenario is that we can query a bag $\mathbf{B}_i$, whose output $f(\mathbf{B}_i)$ (here, this output is obtained by a multiple-instance algorithm) is the closest to 0. The criterion function for $\mathbf{B}_i$ can then be determined as

$$Crit(\mathbf{B}_i) = -f(\mathbf{B}_i)^2 \tag{1}$$

We name the MIAL methods with this criterion function as S-MIAL. It can be seen that the method in [4] is actually a special case of S-MIAL, when the base classifier is updated by MILES. Of course, the base learner can also be generalized to other SVM [32]-based MIL algorithms. such as: *mi-SVM* [12], *MI-SVM* [12], *MI-CA* [33] and *deterministic annealing mi-svm* [13], etc. It is also clear that this method is not restricted to the typical MIL assumption, since we only need the output of $\mathbf{B}_i$ to determine the value of this bag. The whole method can be summarized as in Table 1.

4.2. MIAL with fisher information (F-MIAL)

Although S-MIAL is straightforward and convenient, it neglects the asymmetric information carried by positive and negative bags. This means that, in order to evaluate the value of an unlabeled bag, we should analyze the two cases when it is assigned a positive label and a negative label, respectively. The active learning algorithm presented by Zhang and Oles [28] has given us some inspirations to achieve this goal. In order to devise a classifier, we need to estimate its corresponding parameters $\boldsymbol{\alpha}$. The lower bound for the variance of an estimator t_n for $\boldsymbol{\alpha}$ (here, n means this estimator is based on n samples) can be given by Cramer–Rao inequality [11]. If this estimator is unbiased, the Cramer–Rao inequality can further be simplified to $cov(t_n) \geqslant 1/nI(\alpha)^{-1}$. Here, $I(\alpha)$ is the fisher information.[1] Since it is normally unbiased asymptotically, it is clear that given a classifier, fisher information represents the uncertainty of the estimators for the parameters in this classifier. In [28], for each query, the authors try to find an unlabeled instance that can increase the fisher information most, i.e., decrease the variance most.

Now, let us consider how to use this in MIAL and how to incorporate the asymmetric information carried by positive and negative bags, respectively. Suppose there exists an abstract feature space for bags. Let $q(\mathbf{B})$ be the candidate probability density that can be selected for manual labeling. Denote $I_{q(\mathbf{B})}(\hat{\mathbf{w}})$ as the fisher information matrix of the classification model for distribution $q(\mathbf{B})$ ($\hat{\mathbf{w}}$ is the current estimate of classifier for the t-th query). Here,

$$I_{q(\mathbf{B})}(\hat{\mathbf{w}}) = -\int q(\mathbf{B}) \sum_{y=\pm 1} P(y|\mathbf{B}) \frac{\partial^2}{\partial \mathbf{w}^2} \log P(y|\mathbf{B})\, d\mathbf{B} \tag{2}$$

We try to find a distribution $q(\mathbf{B})$ that can maximize the trace of $I_{q(\mathbf{B})}(\hat{\mathbf{w}})$.[2] For an unlabeled bag $\mathbf{B}_i$, by using sampling techniques [34], Eq. (2) can be transformed to

$$Crit(\mathbf{B}_i) = -\mathrm{tr}\left(\sum_{y_i=\pm 1} P(y_i|\mathbf{B}_i) \frac{\partial^2}{\partial \mathbf{w}^2} \log P(y_i|\mathbf{B}_i) \right) \tag{3}$$

It is clear that, in this equation, we need to calculate second-order derivatives of the conditional probability $P(y_i|\mathbf{B}_i)$, and this will enable us to analyze the conditional probability for instances in unlabeled bags more thoroughly by assigning positive and negative labels to them (we can see this more clearly later in Eq. (5)), which will automatically consider the asymmetric information carried by positive and negative bags. But for S-MIAL, we only care $P(y_i|\mathbf{B}_i)$ ($P(y_i|\mathbf{B}_i)$ is directly related to the $f(\mathbf{B}_i)$ in Eq. (1). We can establish this link simply by using logistic regression), and therefore lose this ability. It is also clear that Eq. (3) is impractical in practice, due to its second-order derivative form. Therefore, we need to get a more concrete expression.

We first assume the base classifier takes the form: $f(\mathbf{B}_{ij}) = \mathbf{w}^T \mathbf{B}_{ij}$. By employing the logistic regression model, we can get: $P(y_{ij}|\mathbf{B}_{ij}, \mathbf{w}) = 1/(1 + \exp(-\mathbf{w^T B}_{ij} y_{ij}))$. To characterize the relationship between $P(y_i|\mathbf{B}_i)$ and $P(y_{ij}|\mathbf{B}_{ij})$,[3] the *noisy-or model* [8][4] is frequently used in multiple-instance learning, and can be expressed as

$$\begin{aligned} P(y_i = +1|\mathbf{B}_i, \mathbf{w}) &= 1 - \prod_j (1 - P(y_{ij} = +1|\mathbf{B}_{ij}, \mathbf{w})) \\ P(y_i = -1|\mathbf{B}_i, \mathbf{w}) &= \prod_j (1 - P(y_{ij} = +1|\mathbf{B}_{ij}, \mathbf{w})) \end{aligned} \tag{4}$$

The form of Eq. (3) can take advantage of the noisy-or model that characterizes the relationship between bags and instances. By employing the logistic regression and substituting Eq. (4) into Eq. (3), we can get a more concrete criterion function:

$$\begin{aligned} Crit(\mathbf{B}_i) = &- \mathrm{tr}\left(\sum_{y_i=\pm 1} P(y_i|\mathbf{B}_i) \frac{\partial^2}{\partial \mathbf{w}^2} \log P(y_i|\mathbf{B}_i) \right) \\ = &\frac{1}{P(y_i = +1|\mathbf{B}_i)} \times (1 - P(y_i = +1|\mathbf{B}_i)) \\ &\times \sum_{j \in \mathbf{B}_i} (P(y_{ij} = +1|\mathbf{B}_{ij}) \times \mathbf{B}_{ij}^T) \sum_{j \in \mathbf{B}_i} (P(y_{ij} = +1|\mathbf{B}_{ij}) \times \mathbf{B}_{ij}) \end{aligned} \tag{5}$$

Here, the index j in the summation of this formula denotes the indices for all the instances in the bag $\mathbf{B}_i$. Since this criterion function

[1] For a detailed definition of the fisher information, refer to [11].

[2] In fact, we try to minimize $\mathrm{tr}(I_{q(\mathbf{B})}(\hat{\mathbf{w}})^{-1} I_{p_u(\mathbf{B})}(\hat{\mathbf{w}}))$ with respect to $q(\mathbf{B})$, where $p_u(\mathbf{B})$ is the distribution of unlabeled bags. But $I_{p_u(\mathbf{B})}(\hat{\mathbf{w}})$ is not a function of $q(\mathbf{B})$. Therefore, it can be neglected in this optimization problem.

[3] We will not distinguish $P(y_i = +1|\mathbf{B}_i, \mathbf{w})$ ($P(y_{ij} = +1|\mathbf{B}_{ij}, \mathbf{w})$) with $P(y_i = +1|\mathbf{B}_i)$ ($P(y_{ij} = +1|\mathbf{B}_{ij})$).

[4] Another commonly used model is the most-likely-cause model [35]. In fact, these two models are quite similar in the performance according to the analysis in [35]. But since the noisy-or model has a nice differentiable property, it is employed here.

utilizes the fisher information matrix, we name it *F-MIAL*. It consists of two terms. The first term $(1/P(y_i{=}{+}1|\mathbf{B}_i))\times(1-P(y_i{=}{+}1|\mathbf{B}_i))$ favors bags with a relatively lower conditional probability $P(y_i{=}{+}1|\mathbf{B}_i)$. The second term can be regarded as the norm of the sum of weighted instances in the bag $\mathbf{B}_i$. The weight for each instance is determined by its conditional probability $P(y_{ij}{=}{+}1|\mathbf{B}_{ij})$. This means that the instance whose positive probability and norm are higher will play a more important role in this term. When these two terms are combined together, it can be deemed as a trade-off between positive instances and negative bags.

In fact, Eq. (5) is a quite general expression. When different MIL methods are used to train the base classifier, it should take different forms. Among all our experiments, we choose mi-SVM [12] (we have given a brief introduction to mi-SVM in Appendix) for simplicity to train the base classifier.[5] Therefore, next we will give a specific expression of Eq. (5) in this case.

The final classifier of mi-SVM [12] can be deemed as the output of an ordinary SVM with the instance label y_{ij} being imputed by some heuristic steps. Therefore, we can analyze its properties from the perspective of an ordinary SVM [36] and the base classifier also takes the form: $f(\mathbf{x})=\sum_i\theta_i\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x})\rangle+b$.[6] Here, θ_i is the dual variable, where the normal vector of the decision boundary is: $\mathbf{w}=\sum_i\theta_i\Phi(\mathbf{x}_i)$. However, criterion equation (5) is based on the assumption that the optimal classifier takes no bias term b. Careful inspection of the ways to solve SVM [36] reveals that b is in fact dependent on θ_i and $\mathbf{x}_i$. Therefore, we try to approximate b in terms of θ_i and $\mathbf{x}_i$.

From the KKT optimality conditions [37] of SVM, all the data points can be attributed into three cases depending on the value of $y_if(\mathbf{x}_i)$, which in turn determines how much the loss is [36]:

1. If $y_if(\mathbf{x}_i)>1$, then $\theta_i=0$.
2. If $y_if(\mathbf{x}_i)=1$, then $\theta_i\in(-C,C)$. (C is the cost weight in SVM.)
3. If $y_if(\mathbf{x}_i)<1$, then $\theta_i=C$ or $-C$.

These three cases refer to instances lying outside, at and inside the margins, respectively. We see that θ_i can take non-extreme values other than 0 and $\pm C$ only if the value of $y_if(\mathbf{x}_i)$ equals 1. The bias b is computed based on the instances at the margins.

We denote the set of instances that lie at the margin for the tth query as $\mathscr{S}(t)$. This set can further be split into two sets, the ones with $y_i{=}{+}1$ and the ones with $y_i{=}{-}1$. These two sets are denoted as $\mathscr{S}^+(t)$ and $\mathscr{S}^-(t)$, respectively. Then, we have the following equations:

$$\sum_i\theta_i\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x})\rangle+b=1,\quad \mathbf{x}\in\mathscr{S}^+(t)$$

$$\sum_i\theta_i\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x})\rangle+b=-1,\quad \mathbf{x}\in\mathscr{S}^-(t)$$

Then, b can be approximated as follows:

$$b=\sum_i\theta_i\times\left\langle\Phi(\mathbf{x}_i),-\frac{\frac{|\mathscr{S}^-(t)|}{|\mathscr{S}^+(t)|}\sum_{j\in\mathscr{S}^+(t)}\Phi(\mathbf{x}_j)+\sum_{k\in\mathscr{S}^-(t)}\Phi(\mathbf{x}_k)}{2|\mathscr{S}^-(t)|}\right\rangle \tag{6}$$

where $|\cdot|$ denotes the number of elements in a set. The solution of SVM can be rewritten as

$$f(\mathbf{x})=\langle\mathbf{w},\Phi(\mathbf{x})-\widehat{\Phi_t(\mathbf{x})}\rangle \tag{7}$$

Table 2
F-MIAL, with mi-SVM as the base MIL learner.

```
Input:
1.  labeled bags: $\{(\mathbf{B_1},y_1),\ldots,(\mathbf{B_{l(0)}},y_{l(0)})\}$, where $l(0)$ is the number of the initially
    labeled bags,
2.  unlabeled bags: $\{\mathbf{B}_{l(0)+1},\ldots,\mathbf{B}_N\}$, where $N$ is the number of bags,
3.  parameters for mi-SVM.
    For $t$:1 to TotalQueryTimes
      ForEach Bag $\mathbf{B}_i$ in the unlabeled set
        Calculate $Crit(\mathbf{B}_i)$ by Eq. (8).
      End ForEach
    Ask query $\mathbf{B_{c(t)}}$ for which $Crit(\mathbf{B}_{c(t)})$ is the largest and get the label $y_{c(t)}$.
    Add $(\mathbf{B}_{c(t)},y_{c(t)})$ to the labeled set, and remove $\mathbf{B}_{c(t)}$ from the unlabeled set.
    Update Classifier $\mathscr{C}$ by mi-SVM.
    End For
Ouput: The final classifier $\mathscr{C}$
```

Here, $\mathbf{w}=\sum_i\theta_i\Phi(\mathbf{x}_i)$, and $\widehat{\Phi_t(\mathbf{x})}$ is defined as $(|\mathscr{S}^-(t)|/|\mathscr{S}^+(t)|)(\sum_{j\in\mathscr{S}^+(t)}\Phi(\mathbf{x}_j)+\sum_{k\in\mathscr{S}^-(t)}\Phi(\mathbf{x}_k))/|2\mathscr{S}^-(t)|$. This amounts to moving the origin of the coordinate to $\widehat{\Phi_t(\mathbf{x})}$ and Eq. (7) can be considered as a linear classifier that contains no bias term. Then we can directly employ the result of Eq. (5).

By substituting $\Phi(\mathbf{x})$ with $\Phi(\mathbf{x})-\widehat{\Phi_t(\mathbf{x})}$, Eq. (5) can be transformed to

$$\begin{aligned}Crit(\mathbf{B}_i)&=\frac{1}{P(Y_i=+1|\mathbf{B}_i)}\times(1-P(Y_i=+1|\mathbf{B}_i))\\&\quad\times\sum_{j\in\mathbf{B}_i}a_j(\Phi(\mathbf{B}_{ij})-\widehat{\Phi_t(\mathbf{x})})^T\times\sum_{j\in\mathbf{B}_i}a_j(\Phi(\mathbf{B}_{ij})-\widehat{\Phi_t(\mathbf{x})})\\&=\frac{1}{P(Y_i=+1|\mathbf{B}_i)}\times(1-P(Y_i=+1|\mathbf{B}_i))\\&\quad\times\left(\sum_{j\in\mathbf{B}_i}a_j^2(1-2E_{S(t),\mathbf{B}_{ij}}+E_{S(t)})\right.\\&\quad\left.+\sum_{j1,j2\in\mathbf{B}_i}a_{j1}a_{j2}(K(\mathbf{B}_{i,j1},\mathbf{B}_{i,j2})-E_{S(t),\mathbf{B}_{i,j1}}-E_{S(t),\mathbf{B}_{i,j2}}+E_{S(t)})\right)\end{aligned} \tag{8}$$

where

$$e=\frac{|\mathscr{S}^-(t)|}{|\mathscr{S}^+(t)|}$$

$$a_j=P(y_{ij}=+1|\mathbf{B}_{ij})=\frac{1}{1+\exp(-f(\mathbf{B}_{ij}))}$$

$$E_{S(t),\mathbf{x}}=\frac{1}{2|\mathscr{S}^-(t)|}\times\left(e\sum_{k\in\mathscr{S}^+(t)}K(\mathbf{x},\mathbf{x}_k)+\sum_{k\in\mathscr{S}^-(t)}K(\mathbf{x},\mathbf{x}_k)\right)$$

$$\begin{aligned}E_{S(t)}&=\frac{1}{4|\mathscr{S}^-(t)|^2}\times\left(e^2\sum_{i,j\in\mathscr{S}^+(t)}K(\mathbf{x}_i,\mathbf{x}_j)+2e\sum_{i\in\mathscr{S}^+(t),j\in\mathscr{S}^-(t)}K(\mathbf{x}_i,\mathbf{x}_j)\right.\\&\quad\left.+\sum_{i,j\in\mathscr{S}^-(t)}K(\mathbf{x}_i,\mathbf{x}_j)\right)\end{aligned}$$

Eq. (8) is the criterion function when mi-SVM is utilized as the base MIL algorithm.

Using mi-SVM as the base learner, F-MIAL can be summarized in Table 2. In fact, the MIL methods used to train the base classifier can also be generalized naturally to other SVM-based MIL algorithms, such as *MI-SVM* [12], *MI-CA* [33] and *deterministic annealing MI-SVM* [13]. Furthermore, to derive this formula, we used the noisy-or model and noisy-or model is a probabilistic model under the

[5] For a detailed description of mi-SVM, refer to Appendix.

[6] $\Phi(\mathbf{x})$ is a linear or nonlinear map, as in [32]. The linear case can be considered as $\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x})\rangle=K_{linear}(x_i,x)=\langle x_i,x\rangle$, and for the nonlinear case, $\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x})\rangle=K_{nonlinear}(x_i,x)$. $K_{nonlinear}$, $K_{nonlinear}$ denote the linear and nonlinear kernels, respectively.

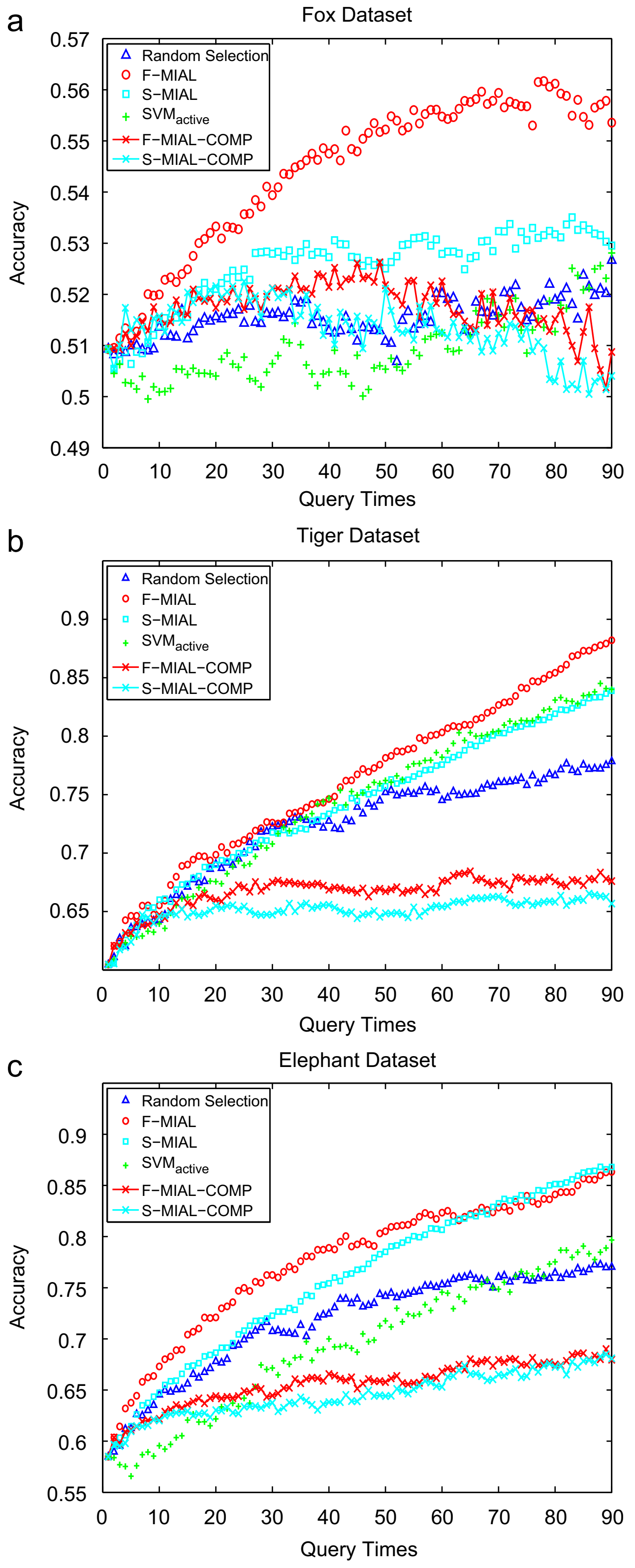

Fig. 1. The comparison results of SVM$_{active}$, S-MIAL, F-MIAL, Random Selection, S-MIAL-COMP and F-MIAL-COMP on Corel dataset.

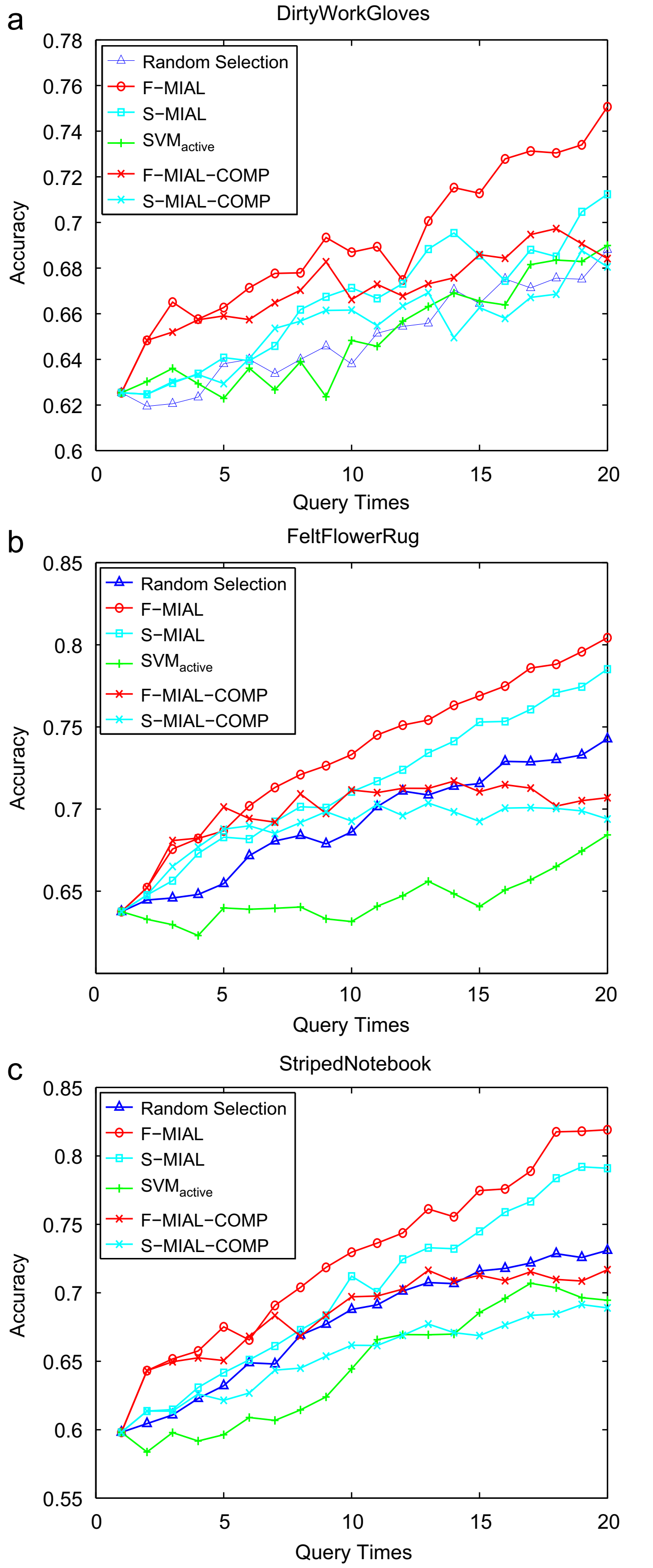

Fig. 2. The comparison results of SVM$_{active}$, S-MIAL, F-MIAL, Random Selection, S-MIAL-COMP and F-MIAL-COMP on SIVAL dataset.

typical MIL assumption, as mentioned before. It is clear that if we change the noisy-or model to the probabilistic models under some other more complicated MIL assumptions. We can extend F-MIAL to those cases.

5. Experiments and discussions

In this section, we report our experimental results in LCBIR. Among all the experiments, we use mi-SVM [12] to train the base

Table 3
The detailed description of the datasets.

Dataset	Features (nonzero)	Bags		Instances		Instances per bag		
		Positive	Negative	Positive	Negative	Min	Max	Ave.
Fox	230 (143)	100	100	647	673	2	13	6.6
Tiger	230 (143)	100	100	544	676	2	13	7.0
Elephant	230 (143)	100	100	762	629	1	13	6.1
DirtyWorkGloves	30 (30)	60	60	1860	1860	31	31	31
FeltFlowerRug	30 (30)	60	60	1860	1860	31	31	31
StripedNotebook	30 (30)	60	60	1860	1860	31	31	31

classifier. We use two baseline methods. The first one is the random selection strategy, where we select an unlabeled bag randomly for labeling for each query. A second baseline is to compute the query quality criteria of S-MIAL and F-MIAL once, and use them throughout the experiment without updating as the classifier changes. In Figs. 1 and 2, F-MIAL-COMP and S-MIAL-COMP correspond to this method when the criterion function of F-MIAL and S-MIAL is computed once at the very beginning. Furthermore, we have also compared the performance of our methods with SVM$_{active}$ by selecting the bag that contains the most uncertain instance for each query. Parameters, such as the bandwidth of the RBF kernel and the cost weight in mi-SVM, are determined by fivefold cross validation.

5.1. Corel dataset

We use three datasets-Fox, Tiger, Elephant[7] from Corel, with the task being to separate these three kinds of animals from other background pictures. The details of this dataset are described in Table 3. For each run, the number of initially labeled bags is set to 20 (10 randomly selected positive bags and 10 randomly selected negative bags). There are 90 queries in total. The average results of 50 independent runs are reported in Fig. 1.

5.2. SIVAL dataset

We also conduct three experiments on SIVAL database.[8] For each experiment, we select all the 60 pictures from a specific category as positive ones, and 60 pictures randomly chosen from other categories as background pictures. The details of these generated datasets are described in Table 3. For each run, the number of initially labeled bags is 10 (five randomly selected positive bags and five randomly selected negative bags). There are 20 queries. The average results of 50 independent runs are reported in Fig. 2.

5.3. Discussions

As can be seen from the experimental results, F-MIAL achieves the best performances in most cases. This is because the second-order derivative form of F-MIAL in Eq. (3) enables us to analyze the conditional probability for instances in unlabeled bags more thoroughly, and therefore considers the asymmetric information carried by negative and positive bags more naturally.

S-MIAL performs well on several datasets. But its performances are worse than F-MIAL. This is due to two reasons. First, the theoretical basis of S-MIAL is the maximum shrinking of the version space. However, this cannot always be ensured due to the limitation of the Simple Margin Strategy [31], especially under the multiple-instance setting where the global optimal solution cannot be ensured. Second, S-MIAL does not consider the asymmetric information carried by positive and negative bags, as we have mentioned previously.

When it comes to SVM$_{active}$, as we can see, although sometimes its performance is comparable with the S-MIAL, it is quite unstable and is much worse than F-MIAL. This is due to the fact that the most uncertain instance can tell little about the positive probability of the corresponding bag.

For S-MIAL-COMP and F-MIAL-COMP, we can see that their performances are worse than S-MIAL and F-MIAL, respectively. This also shows the effectiveness of the interactive scheme under the proposed criterion functions.

As for the time complexity, it is true that the random-selection, S-MIAL, and SVM$_{active}$ are pretty fast and F-MIAL needs some additional calculation. But this extra computational burden is quite trivial, and can be neglected. Because, from Eq. (8), we can see that the calculation burden depends largely on the size of $\mathscr{S}(t)$, which is merely a small part of the support vectors for the tth query. Note that due to the sparsity of the support vectors, the set $\mathscr{S}(t)$ is also very small. Among all the experiments, we find this extra calculation burden negligible.

6. Conclusions and future works

In this paper, to deal with the active learning in LCBIR, we present two general MIAL methods—S-MIAL, which is based on the simple margin strategy, and F-MIAL which is based on the fisher information and the typical MIL assumption. We have shown their superior performances on some real-world datasets, where F-MIAL seems more adapted to the multiple-instance setting than the others. Furthermore, in this paper, our main focus is on the binary classification. However, in LCBIR, the multiple-label problem exists [16]. Therefore, it would be valuable to extend our methods to the multi-label case. And, in active learning, model selection is a very important issue, which enables us to use different models for different queries. We will also consider how to incorporate model selections into our MIAL methods in the future.

Acknowledgements

The work was supported by the National Natural Science Foundation of China under the Grants 60835002, 60721003, and 60605002, and the Research Fund for the Doctoral Program of Higher Education of China under the Grant 200800061056. The authors also wish to gratefully thank all anonymous reviewers who provided insightful and helpful comments.

Appendix

Here, we briefly introduce the basic motivation of mi-SVM [12]. The introduction of SVM will be omitted and the readers could refer

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

[8] http://www.cs.wustl.edu/~sg/multi-inst-data/

to the book [32]. The mixed integer formulation of mi-SVM is [12]

$$\begin{aligned}
\min_{\{y_{ij}\}} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_{ij} \\
\text{s.t.} \quad & \forall i : y_{ij}(\langle \mathbf{w}, \mathbf{B_{ij}} \rangle + b) \geqslant 1 - \xi_{ij}, \quad \xi_{ij} \geqslant 0, \ \ y_{ij} \in \{-1, 1\} \\
& \sum_{j \in B_i} \frac{y_{ij}+1}{2} \geqslant 1, \quad \forall i \text{ s.t. } y_i = 1 \\
& y_{ij} = -1, \quad \forall i \text{ s.t. } y_i = -1
\end{aligned} \tag{9}$$

Here, b is the bias of the classifier. In the traditional single-instance learning problem, the labels y_{ij} for the instance $\mathbf{B_{ij}}$ are known. But in Eq. (9), the labels in the positive bags are treated as unknown binary variables. It is obvious that this is a non-convex optimization problem. Therefore, a heuristic method is employed to solve this problem. Initially, the instance labels are assigned to be the bag labels. Then, a classifier is trained based on these instances and their corresponding pseudo-labels. After that, we update the pseudo-instance labels based on the newly updated classifier. This process will be repeated until no instance label changes any more.

References

[1] D. MacKay, Information-based objective functions for active data selection, Neural Computation 4 (1992) 590–604.

[2] R. Rahmani, S.A. Goldman, H. Zhang, J. Krettek, J.E. Fritts, Localized content based image retrieval, in: MIR'05, ACM Press, New York, NY, USA, 2005, pp. 227–236.

[3] X.S. Zhou, T.S. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimedia Systems 6 (2003) 536–544.

[4] J. Meessen, X. Desurmont, J.-F. Delaigle, C.D. Vleeschouwer, B.M. Macq, Progressive learning for interactive surveillance scenes retrieval, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

[5] Y. Chen, J. Bi, J.Z. Wang, Miles: multiple-instance learning via embedded instance selection, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, Nashville, TN, 2006, pp. 1931–1947.

[6] B. Settles, M. Craven, S. Ray, Multiple instance active learning, in: Advances in Neural Information Processing Systems, 2007.

[7] T.G. Dietterich, R.H. Lathrop, T. Lozano-Perez, Solving the multiple instance problem with axis-parallel rectangles, in: Artificial Intelligence, 1998, pp. 1–8.

[8] O. Maron, T. Lozano-Prez, A framework for multiple-instance learning, Advances in Neural Information Processing Systems, vol. 10, MIT Press, Cambridge, MA, 1998, pp. 570–576.

[9] O. Maron, A.L. Ratan, Multiple-instance learning for natural scene classification, in: International Conference on Machine Learning'98, Morgan Kaufmann, San Francisco, CA, 1998, pp. 341–349.

[10] Q. Zhang, S.A. Goldman, EM-DD: an improved multi-instance learning technique, Advances in Neural Information Processing Systems, vol. 14, MIT Press, Cambridge, MA, 2002, pp. 1073–1080.

[11] G. Casella, R. Berger, Statistical Inference, Duxbury Resource Center, 2001 pp. 335–338.

[12] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2003, pp. 561–568.

[13] P. Gehler, O. Chapelle, Deterministic annealing for multiple-instance learning, in: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics, 2007.

[14] T. Gartner, P.A. Flach, A. Kowalczyk, A.J. Smola, Multi-instance kernels, in: International Conference on Machine Learning'02, 2002, pp. 179–186.

[15] J. Kwok, P.-M. Cheung, Marginalized multi-instance kernels, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2007.

[16] Z.-H. Zhou, M.-L. Zhang, Multi-instance multi-label learning with application to scene classification, in: Advances in Neural Information Processing Systems, 2006.

[17] Z.-H. Zhou, J.M. Xu, On the relation between multi-instance learning and semi-supervised learning, in: International Conference on Machine Learning'07, 2007.

[18] C.B. Razvan, R.J. Mooney, Multiple instance learning for sparse positive bags, in: International Conference on Machine Learning'07, 2007.

[19] Y. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, Journal of Machine Learning Research 5 (2004) 913–939.

[20] G. Schohn, D. Cohn, Less is more: active learning with support vector machines, in: Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000, pp. 839–846.

[21] H.S. Seung, M. Opper, H. Sompolinsky, Query by committee, in: Proceedings of the 5th Workshop on Computational Learning Theory, Morgan Kaufmann, San Mateo, CA, 1992, pp. 287–294.

[22] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, in: International Conference on Machine Learning'00, Morgan Kaufmann, San Francisco, USA, 2000, pp. 999–1006.

[23] Y. Freund, E. Shamir, N. Tishby, Selective sampling using the query by committee algorithm in: Machine Learning, 1997, pp. 133–168.

[24] T. Mitchell, Machine Learning, McGraw-Hill, New York, NY, 1997.

[25] D.A. Cohn, Z. Ghahramani, M.I. Jordan, Active learning with statistical models, Journal of Artifical Intelligence Research 4 (1996) 129–145.

[26] S.C.H. Hoi, R. Jin, J. Zhu, M.R. Lyu, Batch mode active learning and its application to medical image classification, in: ICML'06: Proceedings of the 23rd International Conference on Machine Learning, ACM Press, New York, NY, USA, 2006, pp. 417–424.

[27] K. Yu, J. Bi, V. Tresp, Active learning via transductive experimental design, in: International Conference on Machine Learning'06, ACM Press, New York, NY, USA, 2006, pp. 1081–1088.

[28] T. Zhang, F.J. Oles, A probability analysis on the value of unlabeled data for classification problems, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 1191–1198.

[29] Y. Baram, R. El-Yaniv, K. Luz, Online choice of active learning algorithms, Journal of Machine Learning Research 5 (2004) 255–291.

[30] T. Osugi, D. Kun, S. Scott, Balancing exploration and exploitation: a new algorithm for active machine learning, in: International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2005, pp. 330–337.

[31] S. Tong, Active learning: theory and applications, Ph.D. Thesis, 2001.

[32] B. Scholkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[33] O.L. Mangasarian, E.W. Wild, Multiple instance classification via successive linear programming, Technical Report, Data Mining Institute, University of Wisconsin, 2005.

[34] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer, Berlin, 2001.

[35] O. Maron, Learning from ambiguity, Technical Report AITR-1639, 1998.

[36] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, Berlin, 2006.

[37] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, 2004.

About the Author—DAN ZHANG is now a graduate student in Computer Science Department, Purdue University, West Lafayette, IN, USA. Before this, he got this Master's degree from Department of Automation, Tsinghua University, Beijing, China. His research interests include Information Retrieval, machine learning, data mining and pattern recognition.

About the Author—FEI WANG is currently a Postdoctoral Researcher in School of Computer Science, Florida International University, Miami, FL, USA. Before this, he got his PhD's degree from Department of Automation, Tsinghua University. His main research interests include machine learning, data mining and pattern recognition.

About the Author—ZHENWEI SHI was a Postdoctoral Researcher in the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He is currently an associate professor in the Image Processing Center, School of Astronautics, Beijing University of Aeronautics and Astronautics. His research interests include blind signal processing, image processing, pattern recognition, machine learning and neuroinformatics.

About the Author—CHANGSHUI ZHANG is currently a professor in the Department of Automation, Tsinghua University. He is an Associate Editor of the journal Pattern Recognition. His interests include artificial intelligence, image processing, pattern recognition, machine learning and evolutionary computation, etc.