Information Sciences 244 (2013) 142–160

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

A new and informative active learning approach for support vector machine

Lisha Hu [a,b], Shuxia Lu [a,*], Xizhao Wang [a]

[a] Key Lab. of Machine Learning and Computational Intelligence, College of Mathematics and Computer Science, Hebei University, Baoding 071002, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100864, China

ARTICLE INFO

Article history:
Received 9 August 2011
Received in revised form 24 April 2013
Accepted 13 May 2013
Available online 17 May 2013

Keywords:
Support vector machine
Active learning
Relevance feedback
Diversity

ABSTRACT

Active learning approach has been integrated with support vector machine or other machine-learning techniques in many areas. However, the challenge is: Unlabeled instances are often abundant or easy to obtain, but their labels are expensive and time-consuming to get in general. In spite of this, most existing methods cannot guarantee the usefulness of each query in learning a new classifier. In this paper, we propose a new active learning approach of selecting the most informative query for annotation. Unlabeled instance, which is nearest to the support vector machine's hyperplane learnt from both the unlabeled instance itself and all labeled instances, is selected as the query for annotation. Merits of these queries in learning a new optimal hyperplane have been assured before they are annotated and put into the training set. Experimental results on several UCI datasets have shown the efficiency of our approach.

1. Introduction

Support vector machine (SVM) is proposed by Vapnik in 1995, which is on the basis of the structural risk minimization principle in statistical learning theory [23]. SVM is often used to solve small-sized classification problems and can attain high prediction ability. Suppose there is a pool of unlabeled instances. Instances in the training set of SVM are often selected randomly from the pool and annotated by experts. Such an instance selection method is usually called "the regular passive learning". We can imagine that, if all instances in the training set are selected actively from the pool and are valuable enough, a SVM classifier with higher predictive ability can be learnt on this training set.

With regard to many supervised learning tasks, instances are cheap or free to get, but their labels are difficult or expensive to obtain in general. This issue is usually named as the labeling bottleneck. To overcome this problem, active learning approach offers a way to select valuable instances for annotation on purpose of minimizing indispensable training instances [19]. The main difference between active and passive learning approaches is how queries are selected. Recent works on selecting instances in the training set of SVM mainly rely on active learning approach.

Active learning approach has been applied in areas of text classification [14,22], content-based image retrieval (CBIR) [12,15,21], and so on. One of the central elements in CBIR is the semantic gap between low-level features (color, shape, texture, geometric relationship, etc.) and high-level semantic concepts. As different people may have different visual understandings for the same image, relevance feedback (RF) approach can be used to overcome this gap [16,17,28]. Many approaches have been proposed to incorporate RF with CBIR [1,7,10,20,27].

* Corresponding author. Tel.: +86 1393 0240786; fax: +86 0312 5079638.
E-mail address: cmclusx@126.com (S. Lu).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.05.010

Several approaches have been proposed to perform active learning with SVM [3,5,13,18,22,25]. A frequently-used approach is SVM active learning [22]. It selects each instance from the unlabeled pool on the basis of the version space minimization principle. SVM active learning approach can greatly reduce the demand of labeled instances in areas of text classification [22] and image retrieval [21]. In the batch mode, this approach simply selects a batch of unlabeled instances by the same principle of selecting a single unlabeled instance. However, this process will lead to redundancy in a batch. Fortunately this kind of redundancy can be eliminated by combining diversity measurement with the instance selection approach in the batch mode.

Several kinds of diversity measurement have been proposed by now. For example: With regard to each pair of unlabeled instances in a batch, intersection angles of these instances' corresponding hyperplanes are considered as the degree of angular diversity [2]. Information-theoretic diversity [6] is on the basis of Shannon's entropy and Parzen density estimation. An Active-RDD approach [26]considers the diversity by clustering. Other approaches directly research the batch mode active learning problems in which diversity degree in a batch are considered implicitly [9,11,12,15]. A discriminative batch mode active learning approach [9] considers both the relationship of unlabeled instances and the optimization problem of learning an optimal classifier simultaneously. In a semisupervised SVM batch mode active learning approach [12], either a QP problem or a greedy approach is to be settled in the batch mode.

A new instance selection approach is proposed in this paper, which is on the basis of SVM active learning approach in [22] and SVM batch mode active learning approach in [12]. Procedures of our approach are as follows: An unlabeled instance is put into the training set first. Two optimal hyperplanes corresponding to this instance's two possible labels are learnt from the training set. Then two distance values to these two optimal hyperplanes are calculated for each unlabeled instance. Only the lager distance value is concerned in our approach. So each unlabeled instance is corresponding to a single distance value. An instance, which corresponds to the minimal distance value, is selected finally for annotation. Query of our approach is assured to be the closest to the new hyperplane no matter what its label is. Generally, such queries are support vectors to the new hyperplane. So informativeness of the query can be assured to a great extent in constructing a new classifier. In the batch mode, our approach selects a batch of instances for annotation by leading in the angular diversity in [2]. The reason why we choose the angular diversity rather than other diversity measurements above is that: our approach and the angular diversity have the same theoretical background of version space minimization.

The motivation of our approach is that we want to select such a query, which will be the informative and effective instance in the training set in constructing a new classifier to a great extent. Experimental results on several datasets show the efficiency and effectiveness of the proposed method.

The rest of the paper is structured as follows. Some related work is introduced in Section 2. We present the key idea of our approach and describe the difference between our approach and other approaches in Section 3. Experimental results and discussions of comparing our approach with other two approaches are in Section 4. This paper concludes in Section 5.

2. Related work

2.1. SVM

SVM is proposed on the basis of 2-class classification problem. Here we review the optimization problem of SVM briefly [24]. Let $L = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ be the training set of l instances. $x_i \in R^n$, y_i is the label of x_i. $y_i \in \{-1, 1\}$, $i = 1, \ldots, l$. If the training set is linearly separable, SVM is looking for a hyperplane, which can separate all the instances with a maximal margin. Here the equation of a hyperplane is defined as $f(x) = 0$. Instances in the training set are mapped into a higher dimensional feature space H_k by a mapping ϕ for nonlinear case. The primal optimization problem of SVM is Eq. (1).

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \tfrac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i(\langle\omega, \phi(x_i)\rangle + b) \geqslant 1 - \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geqslant 0, \quad i = 1, \ldots, l \end{aligned} \tag{1}$$

ξ is the slack variable; C is the tradeoff parameter between maximal margin and minimal misclassification. Let the inner product of vectors in the feature space H_k be kernel function k, K is the Gram matrix $K_{ij} = k(x_i, x_j)$. The optimal hyperplane is the Eq. (2).

$$f(x) = \langle\omega^*, \phi(x)\rangle + b^* = 0. \tag{2}$$

ω^* and b^* are the optimal solutions of the problem in (1). We solve the optimization problem in (1) by its dual form and can obtain

$$\begin{aligned} \omega^* &= \sum_{i=1}^{l}\alpha_i^* y_i \phi(x_i), \\ b^* &= y_j - \sum_{i=1}^{l} y_i \alpha_i^* k(x_i, x_j), \quad 0 < \alpha_j < C. \end{aligned} \tag{3}$$

α is the Lagrange multiplier and α^* is the optimal solution of (1)'s dual problem. The decision function is:

$$\operatorname{sgn}(f(x)) = \operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i k(x, x_i) + b^*\right). \tag{4}$$

2.2. Active learning

To overcome the labeling bottleneck mentioned in Section 1, active learning is an approach by selecting informative unlabeled instances for annotation on the purpose of minimizing the labeling times. We summarize the key idea of active learning in [19] and pool-based active learning which can be found in [14].

The core issue in active learning is how to select the next query for annotation. Several strategies have been proposed. For example: Uncertainty sampling approach selects the instance whose label is the most uncertain one to the classifier; Query by committee approach selects the instance whose label is the most divergent for the committees. Other approaches are density-weighted methods, estimated error reduction, and so on. Detailed description of these approaches can be found in [19].

Several scenarios can be found where active learning approach is used essentially to select queries, such as membership query synthesis, stream-based selective sampling, and pool-based active learning [19]. Only Pool-based active learning is reviewed here as it is contained in our approach. Pool-based active learning is as the following. Suppose there are a small-sized dataset L ('labeled set' or 'training set' for short) of labeled instances, and a large-sized dataset U of unlabeled instances ('unlabeled set' for short) which can be considered as a pool of unlabeled instances. In pool based active learning, an informative unlabeled instance x^* is selected from the pool U for annotation, then x^* becomes a labeled instance and is put into the dataset L. Several iterations later, the newly incremental dataset L is conducted as a training set to learn a new classifier for future prediction [14].

2.3. SVM active learning

SVM active learning is proposed in [22] and is recognized as the first technique of combining active learning with SVM. On the premise of all feature vectors' modulus being constant (e.g. $\phi(x) = 1$ for each x) in the feature space H, each feature vector $\phi(x)$ is on the surface of a hyper sphere in H, whose radius is equal to the constant above. Each labeled instance x in the training set L corresponds to a single hyperplane $\langle\omega, \phi(x)\rangle = 0$, which intersects with the hyper sphere through the mapping feature vector $\phi(x)$ of x; each unit normal vector ω of a hyperplane $\langle\omega, \phi(x)\rangle = 0$ in H corresponds to a single point on the sphere. All normal vectors of hyperplanes, which can classify all instances in the training set L correctly, make up the version space. Three approximate approaches of bisecting the version space are proposed in [22]. They are Simple Margin, MaxMin Margin, and Ratio Margin respectively. We briefly review Simple Margin below, as this approach is comparable with the other two approaches in accuracy but more efficient in time.

Firstly Simple Margin approach learns an optimal hyperplane in Eq. (5) from all labeled instances in the training set L.

$$f^*(x) = \langle\omega^*, \phi(x)\rangle + b^* = 0 \tag{5}$$

Then such an unlabeled instance x^*, which is nearest to the hyperplane, is selected from the unlabeled set U for annotation. which is as follows [22]:

$$x^* = \arg\min_{x\in U} \frac{|f^*(x)|}{\|\omega^*\|} \tag{6}$$

Then x^* is removed from the unlabeled set U and put into the labeled set L. Several iterations later, a classifier $\operatorname{sgn}(f^{**}(x))$, which is learnt from the training set L, will be accurate enough to predict all unlabeled instances left in U.

Only one query x^* is found each time in SVM active learning's theory, which is unrealistic in practical application. So authors in [22] select a batch of queries each time with the same principle above. That is, a batch of unlabeled instances, which are nearest to the hyperplane, are selected for annotation. However, experimental results show that such a batch mode instance selection approach leads to redundancy in queries. Fortunately this kind of redundancy can be eliminated by combining diversity measurement with the instance selection approach in the batch mode.

Different kinds of diversity measurements have been introduced in Section 1. Here we briefly review the angular diversity in [2]. We have known that: each instance x corresponds to a hyperplane in the feature space. So the difference between each two instances can be weighed by the diversity of two instances' corresponding hyperplanes. Hyperplanes' diversities can be represented by the cosine value of intersection angles, and can be calculated by the kernel function. Let θ_{ij} be the intersection angle of two hyperplanes corresponding to x_i and x_j. Let k be the kernel function. Then cosine of θ_{ij} is:

$$\cos\theta_{ij} = \left(\frac{|k(x_i, x_j)|}{\sqrt{k(x_i, x_i)k(x_j, x_j)}}\right) \tag{7}$$

So by incorporated with the angular diversity proposed in [2], queries are selected by the incremental strategy below [2]:

$$\begin{aligned} &\textit{batch set } S = \{\ \} \\ &\textit{do}\quad x^* = \arg\min_{x_i \in U-S}\left(\lambda \frac{|f^*(x_i)|}{\|\omega^*\|} + (1-\lambda)\max_{x_j \in S}\quad \cos\theta_{ij}\right) \\ &\qquad S = S \cup \{x^*\} \\ &\textit{until} |S| = \textit{Batchsize} \end{aligned} \tag{8}$$

S is the set of queries which have been selected. f^* is the optimal function learnt on L, that is to say, $f^*(x) = 0$ is the optimal hyperplane learnt on L. *Batchsize* is a predefined parameter which represents the number of instance in a batch. λ is the tradeoff parameter between minimal distance and maximal angle, $\lambda \in [0,1]$.

2.4. Semisupervised SVM batch mode active learning

Semisupervised SVM active learning is proposed in [12] and can be recognized as an improvement of SVM active learning approach in [22]. First a data-dependent kernel function k is learnt by considering the geometric relationship of instances both in the labeled set L and the unlabeled set U. Let ϕ be the mapping function corresponding to the kernel function k. H_k represents the feature space. Then the author in [12] explains that the query's effect in Eq. (6) is overestimated. The reason is: As Eq. (6) is proved to be equivalent to the equation below:

$$x^* = \arg\min_{x \in U} \max_{y \in \{-1,1\}} g(f^*, L \cup \{(x,y)\}, \mathrm{K}) \tag{9}$$

$f^*(x) = \langle \omega^*, \phi(x) \rangle + b^*$ is the optimal function learnt on L; g is the object function in the primal problem of SVM.

$$g(f^*, L, \mathrm{K}) = \frac{1}{2}\|\omega^*\|^2 + C\sum \xi_i^* \tag{10}$$

K represents the Gram matrix of k. x^* in (9) is the most valuable instance on the premise of the hypothesis that: f^* is unchanged even if one more instance (x,y) is put into the training set. So the query's effect in increasing the value of g is overestimated. Then a new instance selection approach in (11) is proposed in [12] to compensate this shortcoming by considering the changes of f when (x,y) is put into the training set.

$$x^* = \arg\min_{x \in U} \max_{y \in \{-1,1\}} \min_{f \in H_k} g(f, L \cup \{(x,y)\}, \mathrm{K}) \tag{11}$$

Eq. (11) is a method of selecting one single query each time in [12]. For semisupervised SVM batch mode active learning, another two approximate methods are proposed in [12], in which a QP problem or a greedy approach is solved in order to approximate to the combinatorial optimization problem for the batch mode of Eq. (11).

3. An informative active learning approach for SVM

3.1. Optimization problem in choosing a single query

In Section 2.4 we describe the weakness of SVM active learning approach. This approach has overestimated each unlabeled instance's importance of learning a classifier. Changes of the hyperplane are ignored when one more instance is annotated and added into the training set. In view of the changes, our approach is proposed below. First of all, we state that query's each possible label does not affect SVM active learning approach's decision. That's the following theorem.

Theorem 1. *Let f^* be the optimal function learnt on labeled instances in the training set L. U is the set of unlabeled instances. Then we have:*

$$x^* = \arg\min_{x \in U} \frac{|f^*(x)|}{\|\omega^*\|} \Longleftrightarrow x^* = \arg\min_{x \in U} \max_{y \in \{-1,1\}} \frac{|f^*(x)|}{\|\omega^*\|} \tag{12}$$

Proof. The equivalence above is easy to prove. For each instance x in the unlabeled set U, the approach in the left part above selects the x^* which can obtain the minimum of $|f^*(x)|/\|\omega^*\|$. When the label y of x is set to be 1 or -1 in the right part, the value $|f^*(x)|/\|\omega^*\|$ in the right part is still equal to the value $|f^*(x)|/\|\omega^*\|$ in the left part no matter what y is equal to. In other words, the value $|f^*(x)|/\|\omega^*\|$ is irrespective of the label y. That is to say, $|f^*(x)|/\|\omega^*\| \Leftrightarrow \max_{y\in\{-1,1\}} |f^*(x)|/\|\omega^*\|$, for the same instance x. □

The motivation of changing Eq. (6) into the right part of Eq. (12) is: we want to prove that: x's each possible label y does not affect the value $|f^*(x)|/\|\omega^*\|$ at all, so y has little effect on SVM active learning's decision of selecting x^* either. Here a new classifier is learnt after x^* is labeled and put into the training set L in our approach.

By considering the changes of f in [12] when (x,y) is added into the training set, the right part of equation above is changed into the equation below:

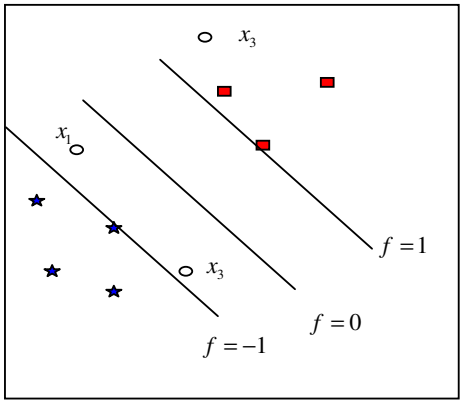

Fig. 1. Initial instances and hyperplane. Red squares and blue stars represent positive and negative instances in the labeled training set L. Circles are unlabeled instances in U. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\begin{aligned} x^* &= \arg \min_{x\in U} \max_{y\in\{-1,1\}} |f^{\#}(x)|/\|\omega^{\#}\| \\ f^{\#} &= \arg \min_{f\in \mathrm{H}_k} g(f, L\cup\{(x,y)\}, \mathrm{K}) \end{aligned} \tag{13}$$

The meaning of notation g is explained in Section 2.4. K represents the Gram matrix. The main idea of our approach in equation (13) is to select such a query x^*, which is nearest to the hyperplane learnt by both instances in L and the query x^* by considering its two possible labels. As is known, training instances around the hyperplane play important roles in constructing an optimal hyperplane. These instances can be support vectors, or misclassified instances by the classifier. So queries should be selected in this important area. However, when a query x^* is labeled and put into the training set L to learn a new hyperplane, this new training instance may be far away from the new hyperplane and become dispensable. That is, we cannot assure that this new training instance is useful to the new hyperplane. However, query selected by Eq. (13) is assured to be closest to the new hyperplane in the worst case of labels. Description of our approach in selecting informative query is shown below.

Initial instances are shown in Fig. 1. The equation of the optimal hyperplane learnt on L is $f=0$. Margin is the distance between $f=1$ and $f=-1$.

For each unlabeled instance x_i, we consider its two possible labels in Fig. 2. For each unlabeled instance x_i, value of distances can reflect the degree of importance in learning a hyperplane. As support vectors except outliers always lie close to the hyperplane, so queries should be close to the hyperplane and be the support vectors as much as possible. Only the larger one between d_i' and d_i'' is concerned in our approach. From Fig. 2 we can see that, d_1'', d_2' and d_3'' are the larger distances corresponding to x_1, x_2 and x_3 respectively. The value of distance concerned in our approach can weigh the degree of importance in x_i's worse case of labels. The final query is the instance whose corresponding larger distance is minimal. A comparison of d_1'', d_2' and d_3'' indicates that x_3 should be selected as the query for future annotation.

x's each possible label plays an important role in the instance selection process of our approach, which is different from SVM active learning. So instance selected by our approach can be as useful as possible in constructing a new hyperplane. The algorithm is presented in Algorithm 1.

Algorithm 1. Our approach in choosing a single query for annotation

Input: small-sized labeled dataset L, large-sized unlabeled dataset U
Output: a unlabeled instance x^* for annotation
Begin
V is set to be a $1\times|U|$ vector of zeros, $|U|$ represents the number of unlabeled instances in U
For each instance x_i in U
  For x_i's each possible label y, $y\in\{-1,1\}$
    train SVM on the dataset $L\cup\{(x,y)\}$ to get the optimal hyperplane $f^{\#}=0$, compute the value of $|f^{\#}(x_i)|/\|\omega^{\#}\|$,
  $V(i)=\max(V(i), |f^{\#}(x_i)|/\|\omega^{\#}\|)$
  End for
End for
find the minimal element min_V in the vector of V, and the ID of min_V,
such that: $V(ID)=$ min_V $x^*=x_{ID}$, $U=U-\{x^*\}$
End

3.2. Optimization problem in choosing a batch of queries

Only one query is selected from the unlabeled pool by using our approach in Section 3.1, which is unrealistic in practical applications. As experts may lose patience after several times of annotation, we discuss how to apply our approach to the batch mode in this section.

We incorporate the angular diversity in [2] with our approach in the batch mode. So a batch of queries is selected by the equation below:

$$\begin{aligned}
&\textit{batch set } S = \{\,\}\\
&\textit{do } \; x^* = \arg\min_{x_i \in U-S} \left(\lambda \max_{y\in\{-1,1\}} \frac{|f^{\#}(x_i)|}{\|\omega^{\#}\|} + (1-\lambda) \max_{x_j \in S} \cos\theta_{ij} \right)\\
&\quad f^{\#} = \arg\min_{f\in H_k} \; g(f, L\cup\{(x_i,y)\}, K)\\
&\quad S = S\cup\{x^*\}\\
&\textit{until} |S| = \textit{Batchsize}
\end{aligned} \tag{14}$$

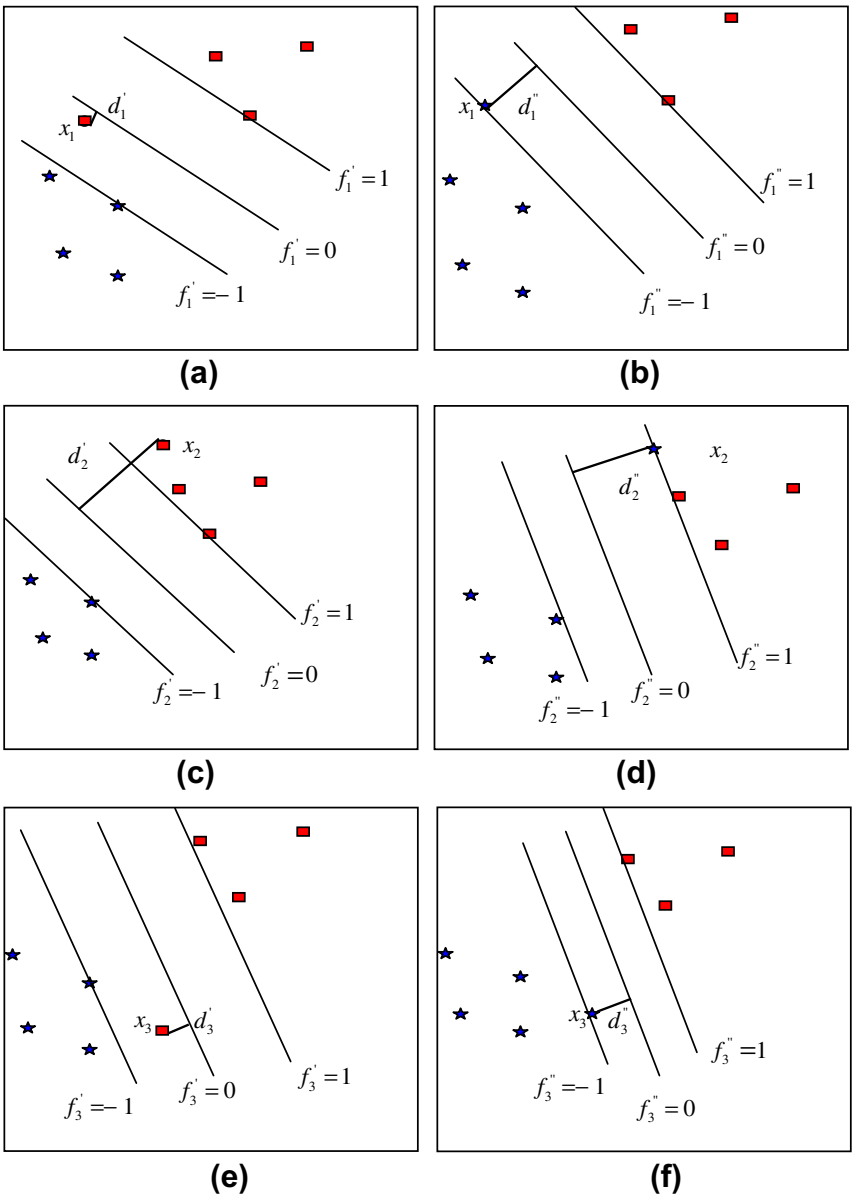

Fig. 2. Distances of unlabeled instances to hyperplanes by considering all possible labels. The label of x_i is taken as positive (or negative) in (a), (c) and (e) (or in (b), (d) and (f)). Hyperplanes $f_i' = 0$ and $f_i'' = 0$ are learnt from L and x_i. d_i' and d_i'' are distances between x_i and these two hyperplanes.

Notations S, k, and *Batchsize* are explained in Section 2.3. The reason we choose angular diversity instead of others is that the angular diversity has the same theory foundation with SVM active learning and the same principle of version space minimization. An instance, which is nearest to the new hyperplane in the worst case of labels, is selected for annotation. Such an instance can approximately bisect the version space. So our approach is also based on the version space minimization principle. We believe the batch mode of our approach can obtain a high performance by incorporating with the angular diversity. The algorithm is presented in Algorithm 2.

Algorithm 2. Our approach in choosing a batch of queries for annotation

Input: small-sized labeled dataset L, large-sized unlabeled dataset U, trade-off parameter λ, number of instances in the batch Batchsize
Output: a batch of unlabeled instances in S for annotation
Begin
V is set to be a $1 \times |U|$ vector of zeros, $|U|$ represents the number of unlabeled instances; $S = \{\}$; $M = 0$
Do
  For each instance x_i in the dataset $U - S$
    For x_i's each possible label y in $\{-1, 1\}$
      train SVM on dataset $L \cup \{(x, y)\}$ to get the optimal hyperplane $f^{\#} = 0$, compute the value of $|f^{\#}(x_i)|/\|\omega^{\#}\|$, $V(i) = \max(V(i), \lambda|f^{\#}(x_i)|/\|\omega^{\#}\|)$
    End for
    For each instance x_j in S
      compute $k1 = |k(x_i, x_j)| \Big/ \sqrt{k(x_i, x_i)k(x_j, x_j)}$, $M = \max(M, k1)$
    End for
    $V(i) = V(i) + (1 - \lambda)M$
  End for
  find the minimal element min_V in the vector of V, and the ID of min_V, such that: $V(ID) = \min\_V$
  $S = S \cup \{x_{ID}\}$, $U = U - \{x_{ID}\}$
Until $|S| = Batchsize$
End

3.3. Difference between our approach and the approach in [12]

Our approach in Eq. (13) is a little similar to Eq. (11) approach proposed in [12]. Below, we will explain that these two approaches are different.

$$x^* = \arg \min_{x \in U} \max_{y \in \{-1,1\}} \min_{f \in \mathrm{H}_k} g(f, L \cup \{(x, y)\}, \mathrm{K}) \tag{11}$$

$$\begin{aligned} x^* &= \arg \min_{x \in U} \max_{y \in \{-1,1\}} |f^{\#}(x)|/\|\omega^{\#}\| \\ f^{\#} &= \arg \min_{f \in \mathrm{H}_k} g(f, L \cup \{(x, y)\}, \mathrm{K}) \end{aligned} \tag{13}$$

All the notations are explained again. L represents the set of labeled instances; U represents the set of unlabeled instances; K represents the Gram matrix; H_k represents the feature space; f represents the optimal function in (15). ϕ represents the kernel mapping function; g represents the objective function in the primal problem of SVM, which can be found in Eq. (10).

$$f = f(x) = \langle \omega, \ \phi(x) \rangle + b \tag{15}$$

The process of Eq. (11) is as the following: First, let a random instance x' in U be fixed; Second, let x''s random label y' from $\{-1, 1\}$ be fixed. Each function f in H_k corresponds to a value of $g(f, L \cup \{(x', y')\}, \mathrm{K})$. The approach in Eq. (11) considers the minimal value $g(f^{\#}, L \cup \{(x', y')\}, \mathrm{K})$ in the set of values $\{g(f, L \cup \{(x', y')\}, \mathrm{K}) | f \in \mathrm{H}_k\}$. g represents the objective function in the primal problem of SVM, so $f^{\#}(x)$ is SVM's optimal function learnt on the training set of $L \cup \{(x', y')\}$. The formula of g can be seen in Eq. (16). This approach may discard many informative instances because of the misclassified error of instances in the training set. Besides, there exists a tradeoff parameter in this approach, which will be tough to control.

The process of our approach is as follows: First, let a random instance x' in U be fixed; Second, let x''s random label y' from $\{-1, 1\}$ be fixed. A function $f^{\#}$ is learnt in H_k whose corresponding value $g(f^{\#}, L \cup \{(x', y')\}, \mathrm{K})$ is minimal in the set $\{g(f, L \cup \{(x', y')\}, \mathrm{K}) | f \in \mathrm{H}_k\}$. So $f^{\#}$ is SVM's optimal function learnt on the training set of $L \cup \{(x', y')\}$. Our approach considers the value in Eq. (17), which is the distance between the instance x' and the hyperplane $f^{\#} = 0$.

These two approaches consider different parts of the training process on the training set $L \cup \{(x', y')\}$: Eq. (11) approach considers the objective function value $g(f^{\#}, L \cup \{(x', y')\}, K)$ in SVM's primal problem; while our approach considers the distance between the instance x' and the hyperplane $f^{\#}(x) = \langle \omega^{\#}, \phi(x) \rangle + b^{\#} = 0$. Our approach can also overcome some drawbacks of the approach in [12]: There are none free parameters to be control, and the informativeness of queries can be assured in our instance selection approach.

$$g(f^{\#}, L \cup \{(x', y')\}, K) = \frac{1}{2}\|\omega^{\#}\|^2 + C\left(\sum_{x_i \in L} \xi_i + \xi'\right) \tag{16}$$

$$\frac{|f^{\#}(x')|}{\|\omega^{\#}\|} = \frac{|\langle \omega^{\#}, \phi(x') \rangle + b^{\#}|}{\|\omega^{\#}\|} \tag{17}$$

$\|\omega^{\#}\|^2$ represents the square of margin's reciprocal; $\sum_{x_i \in L} \xi_i$ represents the misclassified error of instances in L; ξ' represents the misclassified error of x'; C represents the tradeoff parameter between maximal margin and minimal errors.

4. Experimental results and discussion

4.1. Experiments on an artificial dataset by selecting a single query each time

Two datasets are generated from two Gaussian distributions centered at $(-8, -8)$ and $(8, 8)$ with the same variance 8. Each dataset contains 500 instances. These two datasets compose the positive and negative classes which are plotted in Fig. 3. Squares represent positive instances, and stars represent negative instances. So the whole dataset contains 1000 instances in all. We select 0.2% instances (2 instances, 1 positive and 1 negative) randomly from the whole dataset and put them into the labeled training set L(see Fig. 4). For each of the experiments below, we assure that the labeled set L always contains at least one positive and one negative instances. The other 99.8% instances of the whole dataset are put into unlabeled set U. We select one query each time for annotation by applying three approaches: SVM active learning, Eq. (11)(we represent the approach in [12] by 'Eq. (11)' for simplicity), and Our approach. Each query is annotated and put into the labeled set L, and eliminated from U. We do this 20 times for each of the three approaches. For all three approaches, the value of C is fixed to 1000, Gaussian kernel ($k(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/2\sigma^2)$) is used as the kernel function, the value of σ is fixed to 1. LIBSVM [4] toolbox is applied in all the experiments. Results are shown in Figs. 5–7.

We can see that from Figs. 5–7, instances selected by SVM active learning and Eq. (11) are nearly the same in distribution. These instances all lie around the centers tightly. However, instances selected by our approach lie incompactly in the plane. This incompact distribution might represent a higher degree of diversity in between instances selected to a certain extent. In this view, the result of our approach is a little better than that of the other two approaches.

4.2. Experiments on three UCI datasets by selecting a single query each time

Three UCI datasets: Dermatology, Ionosphere, and Thyroid Disease are selected from [8] and each dataset is used separately as the whole dataset in the experiments. Data description of the three datasets is shown in Table 1. Just like

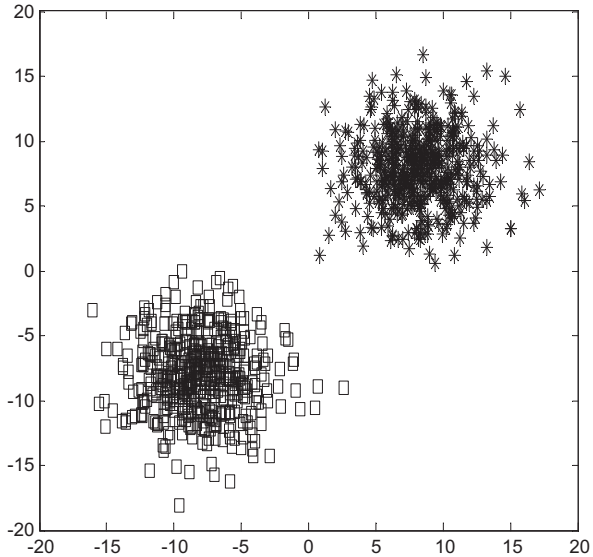

Fig. 3. The whole dataset.

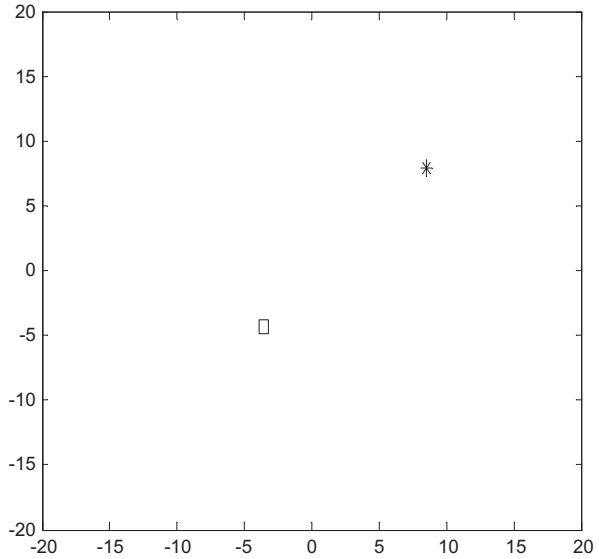

Fig. 4. Initial labeled instances.

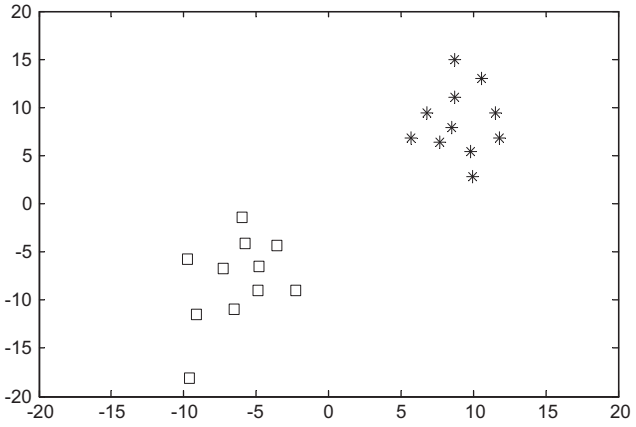

Fig. 5. First 20 instances selected by SVM active learning approach.

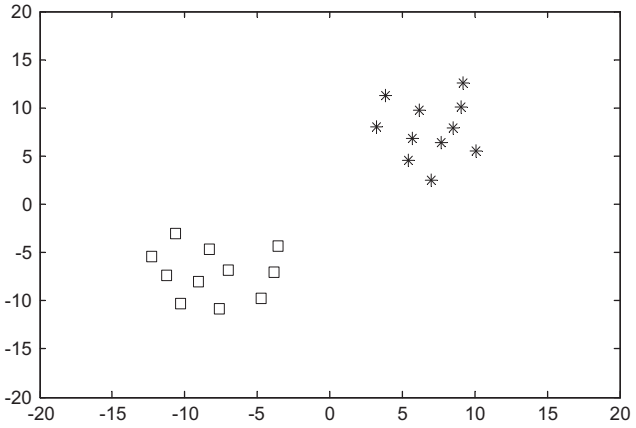

Fig. 6. First 20 instances selected by Eq. (11) approach.

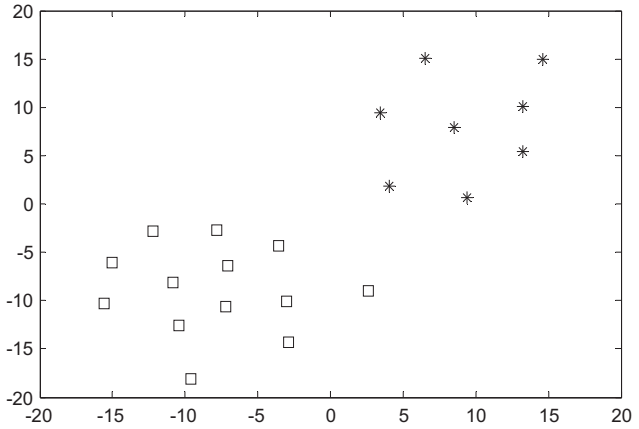

Fig. 7. First 20 instances selected by our approach.

Table 1
Information of three UCI datasets.

Dataset	Sum of positive	Sum of negative	Total number	Dimensionality
Dermatology	171	187	358	34
Ionosphere	224	126	350	34
Thyroid disease	150	65	215	5

the experiment above, we select one query each time for annotation by using the three approaches, and parameters are set the same values as above except $\sigma = 5$ for all the approaches. The reason why we do the experiments here is that we want to compare the abilities of three approaches in choosing the informative unlabeled queries. Test accuracy is calculated for all the approaches on the remaining sets of unlabeled instances. All the experiments are repeated ten times and the results are averaged. Experimental results are shown in Figs. 8–10.

From Figs. 8–10 we can see that, taking as a whole, test accuracy goes up along with the number of queries for most of the three approaches in the three datasets. In Dermatology dataset, our approach and Eq. (11) approach are nearly the same in performance, but our approach can get the highest test accuracy in most of the time, SVM active learning is a little lower in Dermatology; in Ionosphere dataset, our approach performs well most of the time, our approach and Eq. (11) approach can get the highest test accuracy in the end; in Thyroid Disease dataset, although all the three approaches get nearly the same accuracy in the beginning, the accuracy of our approach is a little higher than the other two approaches at last. Eq. (11) approach and SVM active learning are nearly the same in Thyroid.

Besides, we also use correlation coefficient (*cc*) to measure the performance of the three approaches. This indicator can measure the correlation between the forecast result and the actual situation. *cc* is calculated in Eq. (18). All the experiments are repeated ten times and the results are averaged. Experimental results are shown in Figs. 11–13.

$$cc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{18}$$

TP (*or FP*) is the number of positive (or negative) instances which are predicted as positive by the classifier; *TN*(*or FN*) is the number of negative (or positive) instances which are predicted as negative by the classifier. The value of cc ranges from −1 to 1, where 1 represents the forecast result is the same as the actual result and 0 represents the forecast result is a random prediction.

By comparing Figs. 8 with 11, Figs. 9 with 12, and Figs. 10–13 we can see that test accuracy and cc go up and down in the same way, which means that test accuracy can be a valuable indicator for measuring the performance of the three approaches. From Figs. 11–13 we can see that, forecast results of all the three approaches are much better in Dermatology than in other two datasets. Our approach can get the highest prediction results in all the three datasets.

4.3. Experiments on two image datasets by selecting a batch of queries each time

Another two UCI datasets: Image Segmentation and Letter Recognition are also selected from [8] to do the experiments. The Image Segmentation dataset is composed of two sets: the training set and the testing set. Each of these two sets contains 7 classes of instances and each class contains 30 instances in the training set and 300 instances in the testing set. We merge three classes as the positive class, and other four classes as the negative class. The Letter Recognition dataset contains 26

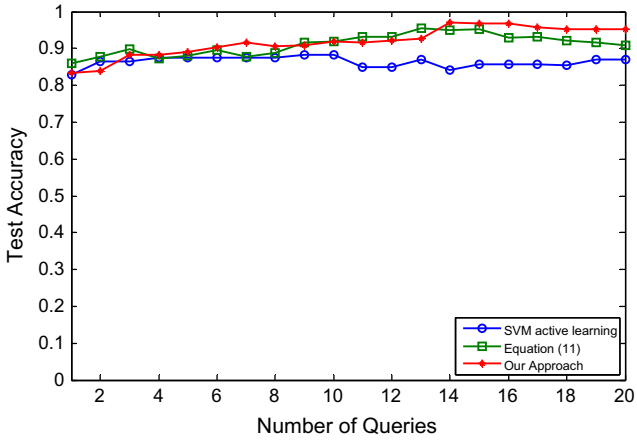

Fig. 8. Test accuracy on Dermatology, first 20 queries is selected.

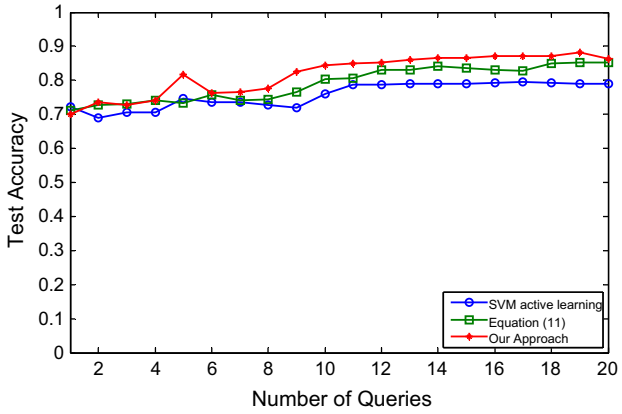

Fig. 9. Test accuracy on Ionosphere, first 20 queries is selected.

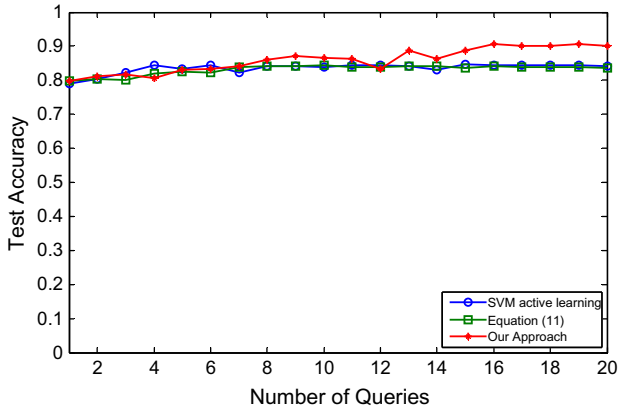

Fig. 10. Test accuracy on Thyroid, first 20 queries is selected.

classes of instances corresponding to 26 letters in the alphabet. Each class contains at least 700 instances. We merge class 'a' and 'b' as positive class, class 'c' and 'd' as negative class. Other classes are eliminated. Then the whole dataset is split randomly into the training set and the testing set of the same size. So both the training set and testing set are changed into 2-class datasets. Description of datasets is shown in Table 2.

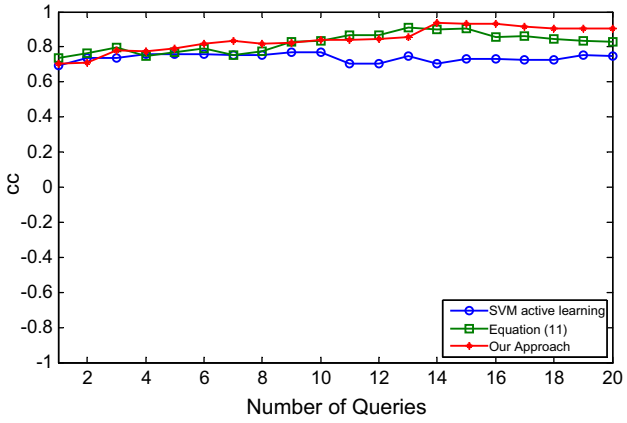

Fig. 11. cc on Dermatology, first 20 queries is selected.

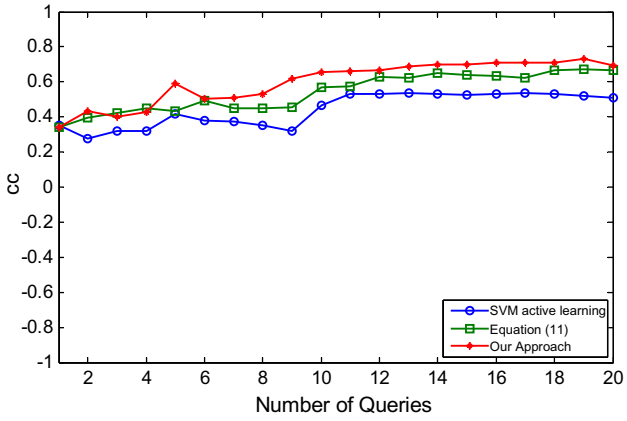

Fig. 12. cc on Ionosphere, first 20 queries is selected.

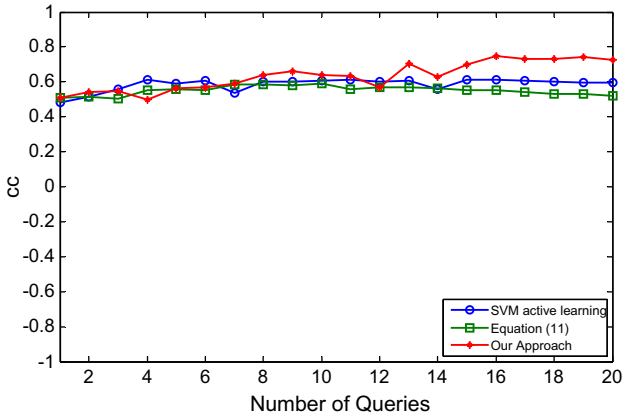

Fig. 13. cc on Thyroid, first 20 queries is selected.

We select 2% and 0.2% instances from the training set as the labeled instances for each image dataset, all other instances in the training set are seen as the unlabeled instances. As seen above, the labeled set contains at least one positive and one negative instance. A batch of queries is selected by the three approaches above. *Batchsize* is fixed to 5, and the value of σ is fixed to 1. Both the SVM active learning approach and our approach have the parameter λ, the range of its value is set to be ten numbers of 0.1, 0.2, ..., 0.9 and 1. The optimal value of λ is the one by using which the approach can attain the highest

Table 2
Information of two UCI datasets.

Dataset	Sum of positive	Sum of negative	Total number	Dimensionality
Image training set	90	120	210	19
Image testing set	900	1200	2100	19
Letter training set	777	771	1548	16
Letter testing set	778	770	1548	16

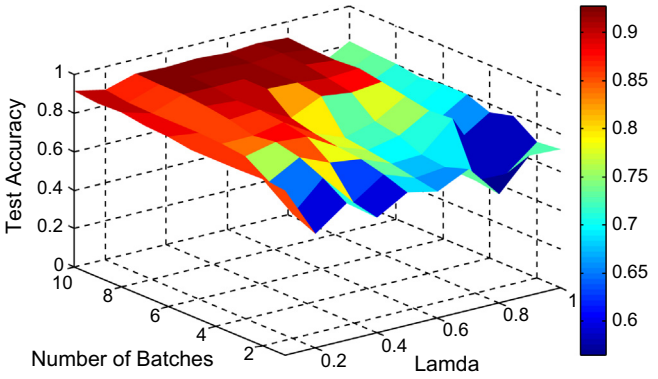

Fig. 14. Test accuracy of SVM active learning approach on Image.

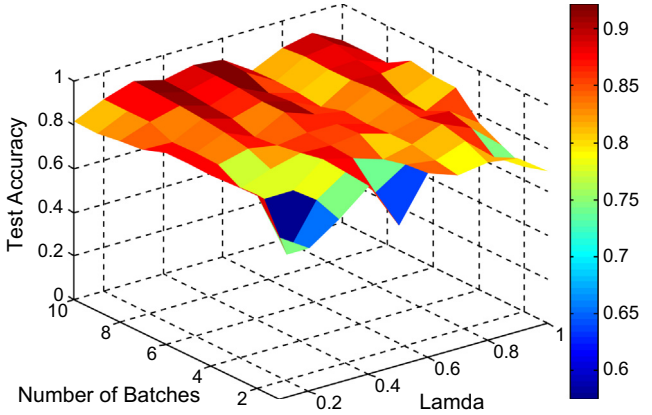

Fig. 15. Test accuracy of our approach on Image.

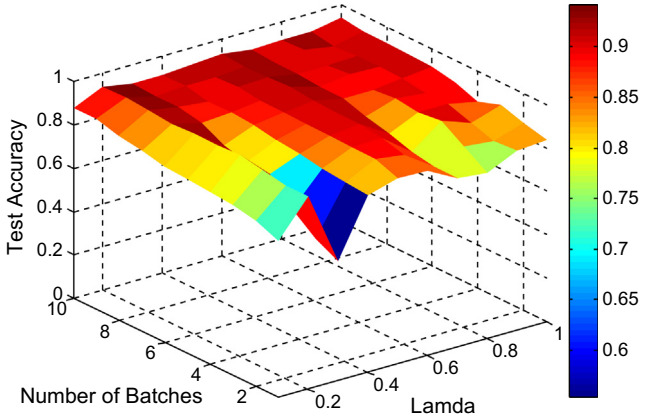

Fig. 16. Test accuracy of SVM active learning approach on Letter.

test accuracy in the testing set after ten batches of queries are annotated and put into the labeled set. Test accuracies of SVM active learning approach and our approach on Image and Letter testing set are shown in Figs. 14–17, respectively.

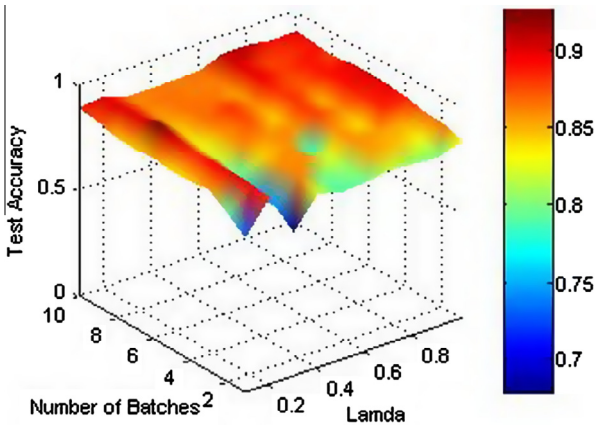

Fig. 17. Test accuracy of our approach on Letter

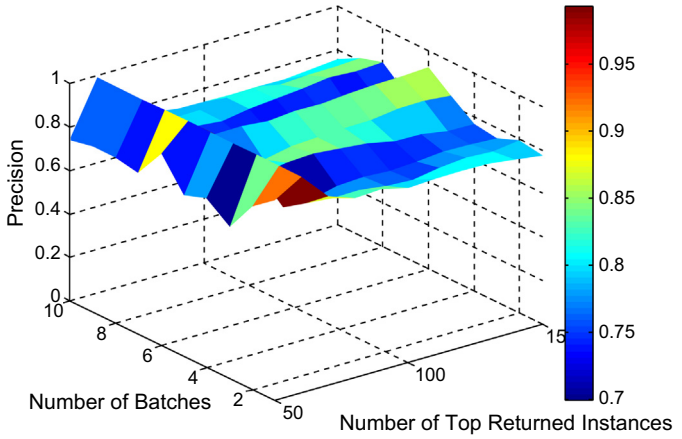

Fig. 18. Precision of SVM active learning approach on Image.

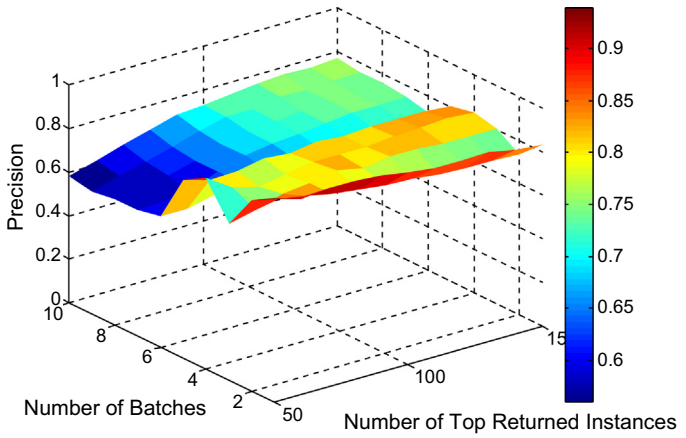

Fig. 19. Precision of Eq. (11) approach on Image.

From the results of Image Segmentation dataset in Figs. 14 and 15, the optimal λ can be either 0.3 or 0.7 for SVM active learning approach, whereas it can be either 0.3 or 0.5 for our approach. So we set λ to be 0.3 for both the two approaches.

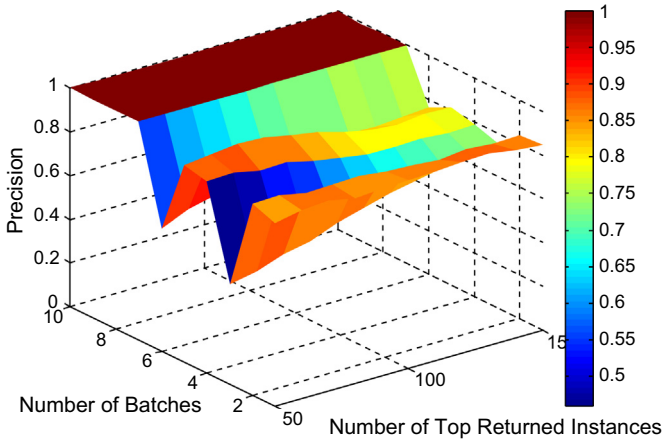

Fig. 20. Precision of our approach on Image.

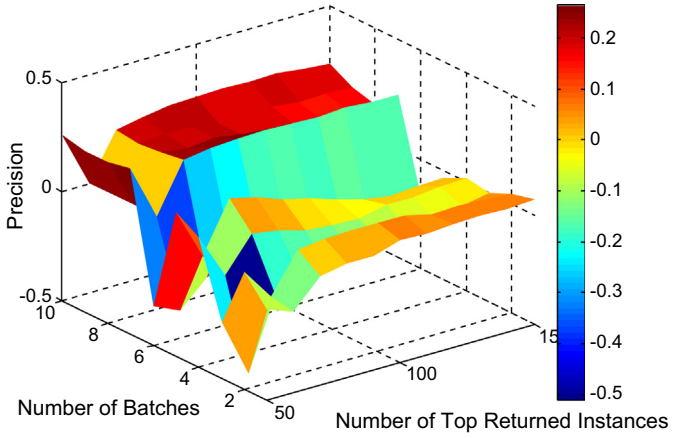

Fig. 21. Precision of our approach minus precision of SVM active learning approach.

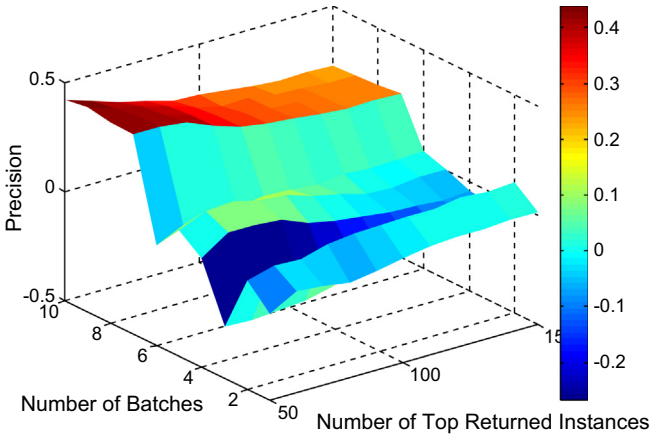

Fig. 22. Precision of our approach minus precision of Eq. (11) approach.

From the results of Letter Recognition dataset in Figs. 16 and 17, the optimal λ is 0.2 for SVM active learning approach, and it can be either 0.7 or 0.8 for our approach. So we set them to be 0.2 and 0.7, respectively.

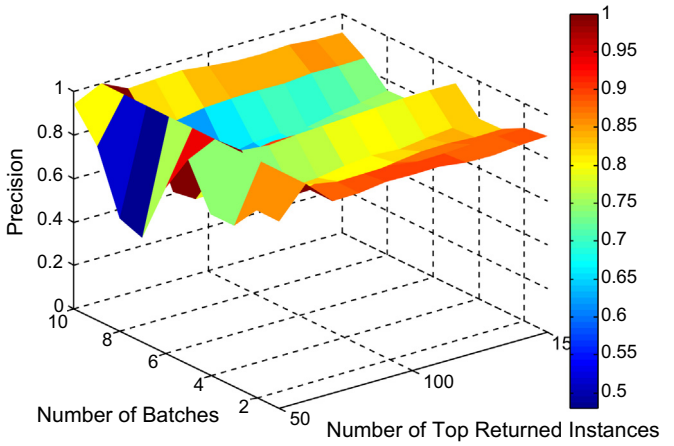

Fig. 23. Precision of SVM active learning approach on Letter.

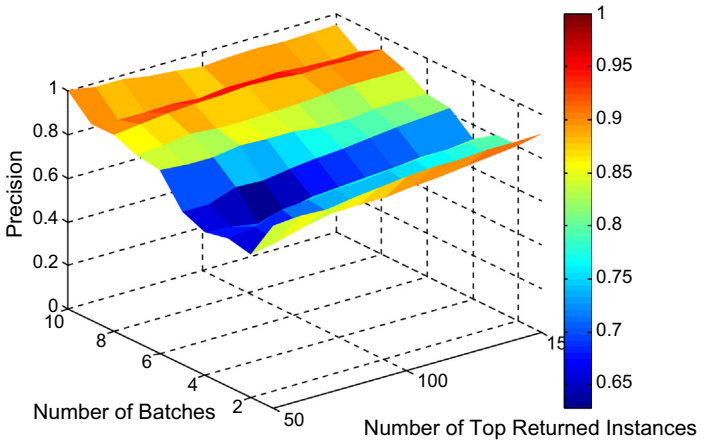

Fig. 24. Precision of Eq. (11) approach on Letter.

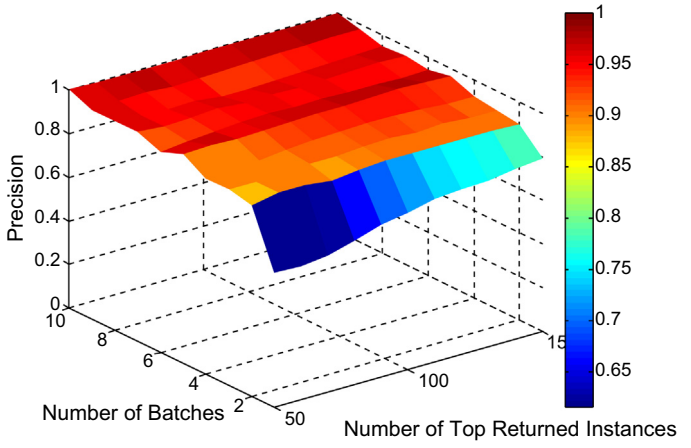

Fig. 25. Precision of our approach on Letter.

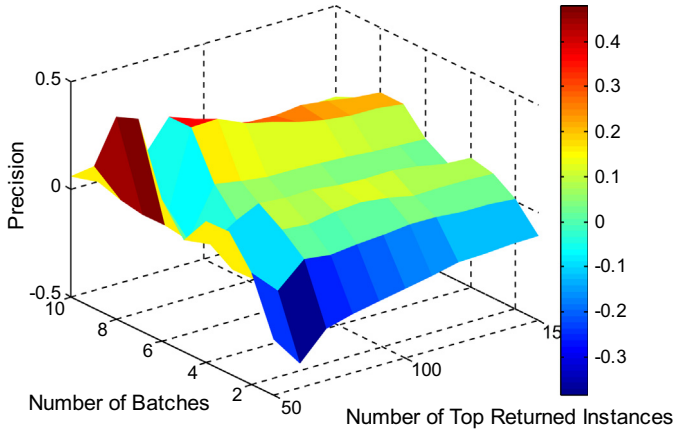

Fig. 26. Precision of our approach minus precision of SVM active learning approach.

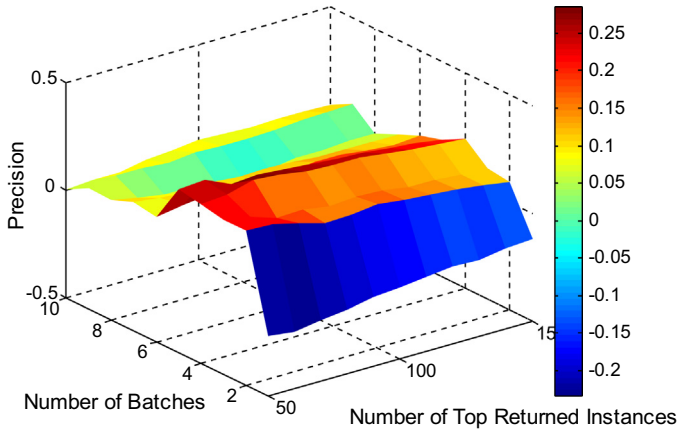

Fig. 27. Precision of our approach minus precision of Eq. (11) approach.

After each parameter λ of the two approaches is set to be the optimal value, these two approaches and Eq. (11) approach are compared in the precision of the testing set. Precision is the percentage of instances whose real labels are positive in instances whose labels the classifier believes are positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

TP (*or FP*) is the number of positive (or negative) instances which are predicted as positive by the classifier. So the experiment process is below: At each time a batch of queries selected by each of the three approaches are annotated and put into the labeled set, which is applied as the training set to learn a new classifier. Then the new classifier predicts instances in the testing set. 50–150 Unlabeled instances which are the most distant from the hyperplane are selected as the top returned instances. These instances are used to compute the precision. Experimental results of the three approaches on the two datasets are shown in Figs. 18–27, respectively.

From the results of those three approaches in Figs. 18–20 we can see that: let the number of top returned instances be fixed, precision of all the three approaches goes up and down when the number of batches is increasing. In other words, precisions of all the three approaches are changing unsteadily. Similar results can also be seen in Figs. 23–25. However, taken as a whole, precision of our approach is increasing, whereas precision of SVM active learning approach only changes a little and precision of Eq. (11) even goes down in the last batch. This means that our approach is a little better than the other two in precision when the number of instances in the labeled training set is growing. Results in Figs. 21 and 22 show how much our approach is higher than the other two approaches in precision. We can see that the values in Figs. 21 and 22 are larger than zero in most cases.

From the results of three approaches in Figs. 23–25 we can see that: all the three approaches' precisions are increasing on the whole. Precision of SVM active learning approach is a little inferior to the other two approaches in Fig. 26. Our approach

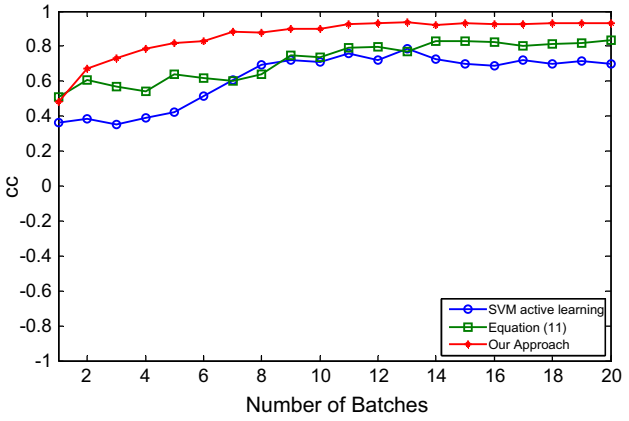

Fig. 28. cc on Image, first 20 batches is selected.

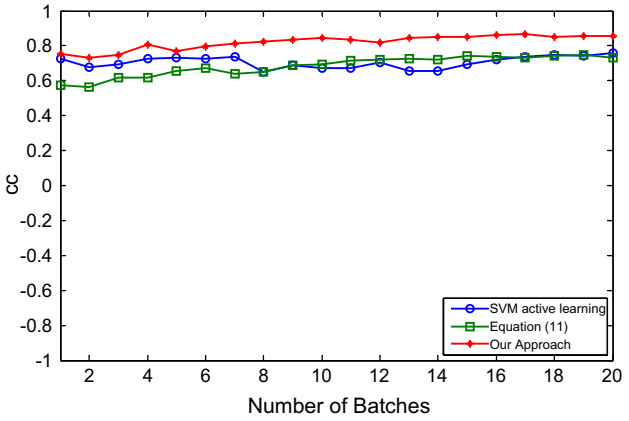

Fig. 29. cc on Letter, first 20 batches is selected.

is a little better than Eq. (11) in the beginning which can be seen in Fig. 27. Eq. (11) approach and our approach are in agreement in precision in the end.

Besides, we also use cc to measure the performance of the three approaches on Image and Letter dataset. Experimental results are shown in Figs. 28 and 29.

From Figs. 28 and 29 we can see that, our approach has better forecast ability on Image and Letter datasets when compared with other two approaches. Especially on Image dataset, the forecast result of our approach can approach to the actual result as much as possible. Forecast results of Eq. (11) approach and SVM active learning approach are nearly the same on Letter dataset, and prediction of our approach is a little better than these two approaches.

5. Conclusion

A new active learning approach for SVM is proposed in this paper. This approach aims at selecting the most informative instance from an unlabeled pool, which is nearest to the new hyperplane learnt from labeled training set and the instance itself. In our approach, a batch of unlabeled instances are selected by considering both the distance and the angular diversity in the batch.

Contributions of our proposed approach are: queries selected by our approach can be better used in learning an optimal hyperplane compared with SVM active learning approach and the approach in [12]. We guarantee that these queries are informative and valuable as they are near the final hyperplane as much as they could.

Our approach is compared with SVM active learning approach and Eq. (11) approach both in selecting a single query and a batch of queries in the experiment. The ways in which our approach and the other two approaches select query from an artificial dataset are also shown in the experiment. Experimental results show the efficiency improvements of our approach.

Acknowledgments

The authors thank the editors and anonymous reviewers. Their valuable and constructive comments and suggestions helped them in significantly improving this paper. This research is supported by the National Natural Science Foundation of China (61170040 and 60903089), by the Natural Science Foundation of Hebei Province (F2011201063 and F2012201023), and by the Key Scientific Research Foundation of Education Department of Hebei Province (ZD2010139).

References

[1] M. Arevalillo-Herráez, F.J. Ferri, J. Domingo, A naive relevance feedback model for content-based image retrieval using multiple similarity measures, Pattern Recognition 43 (2010) 619–629.
[2] K. Brinker, Incorporating diversity in active learning with support vector machines, in: Proceedings of the Twentieth International Conference on Machine Learning, 2003, pp. 59–66.
[3] C. Campbell, N. Cristianini, A. Smola, Query learning with large margin classifiers, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufman, Stanford, CA, USA, 2000, pp. 111–118.
[4] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1–27. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>..
[5] B.J. Cui, H.F. Lin, Z.H. Yang, Uncertainty sampling-based active learning for protein–protein interaction extraction from biomedical literature, Expert Systems with Applications 36 (2009) 10344–10350.
[6] C.K. Dagli, S. Rajaram, T.S. Huang, Leveraging active learning for relevance feedback using an information theoretic diversity measure, in: Proceedings of ACM Conference on Image and Video Retrieval, 2006, pp. 123–132.
[7] C.D. Ferreira, J.A. Santos, R.d.S. Torres, M.A. Gonçalves, R.C. Rezende, W.G. Fan, Relevance feedback based on genetic programming for image retrieval, Pattern Recognition Letters 32 (2011) 27–37.
[8] A. Frank, A. Asuncion, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2010. <http://archive.ics.uci.edu/ml>.
[9] Y.H. Guo, D. Schuurmans, Discriminative batch mode active learning, in: Proceedings of Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2008, pp. 593–600.
[10] D.Q. He, D. Wu, Enhancing query translation with relevance feedback in Translingual information retrieval, Information Processing and Management 47 (2011) 1–17.
[11] S.C.H. Hoi, R. Jin, M.R. Lyu, Large-scale text categorization by batch mode active learning, in: Proceedings of the International Conference on the World Wide Web, Edinbergh, Scotland, 2006, pp. 633–642.
[12] S.C.H. Hoi, R. Jin, J. Zhu, Semisupervised SVM batch mode active learning with applications to image retrieval, ACM Transactions on Information Systems 27 (2009) 1–19.
[13] B. Leng, Z. Qin, L.Q. Li, Support vector machine active learning for 3D model retrieval, Journal of Zhejiang University Science A 8 (2007) 1953–1961.
[14] D.D. Lewis, W.A. Gale, A sequential algorithm for training text classifiers, in: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 13–19.
[15] R.J. Liu, Y.H. Wang, T. Baba, D. Masumoto, S. Nagata, SVM-based active feedback in image retrieval using clustering and unlabeled data, Pattern Recognition 41 (2008) 2645–2655.
[16] Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra, Relevance feedback: a power tool in interactive content-based image relevance, IEEE Transactions on Circuits and Video Technology 8 (1998) 644–655.
[17] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science 41 (1990) 288–297.
[18] G. Schohn, D. Cohn, Less is more: active learning with support vector machines, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufman, Stanford, CA, USA, 2000, pp. 839–846.
[19] B. Settles, Active Learning Literature Survey, Technical Report 1648, Department of Computer Sciences, University of Wisconsin-Madison, Wisconsin, WI, 2009.
[20] D.C. Tao, X.O. Tang, X.L. Li, X.D. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1088–1099.
[21] S. Tong, E. Chang, Support vector machine active learning for image retrieval, in: Proceedings of the ACM International Conference on Multimedia, 2001, pp. 107–118.
[22] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research 2 (2001) 45–66.
[23] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[24] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[25] Z. Wang, S.C. Yan, C.S. Zhang, Active learning with adaptive regularization, Pattern Recognition 4 (2011) 2375–2383.
[26] Z. Xu, R. Akella, Y. Zhang, Incorporating diversity and density in active learning for relevance feedback, in: Proceedings of the European Conference on Information Retrieval Research, SpringerVerlag, 2007, pp. 246–257.
[27] P.Y. Yin, C.W. Liu, A new relevance feedback technique for iconic image retrieval based on spatial relationships, Journal of Systems and Software 82 (2009) 685–696.
[28] X.S. Zhou, T.S. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimedia System 8 (2003) 536–544.