SCIENCE CHINA
Information Sciences

November 2020, Vol. 63 210205:1–210205:15
https://doi.org/10.1007/s11432-020-3082-9

• RESEARCH PAPER •
Special Focus on Multi-source Information Fusion

Active learning based on belief functions

Shixing ZHANG[1], Deqiang HAN[1*] & Yi YANG[2]

[1] *Institute of Integrated Automation, Xi'an Jiaotong University, Xi'an* 710049*, China;*
[2] *School of Aerospace, Xi'an Jiaotong University, Xi'an* 710049*, China*

Received 27 June 2020/Revised 31 August 2020/Accepted 10 September 2020/Published online 22 October 2020

Abstract Active learning involves selecting a few critical unlabeled samples for manual and credible labeling to improve the performance of the current classifier. The critical step of active learning is the sample selection strategy. Uncertainty sampling is a well-known sample selection strategy, which involves selecting the samples for which the current classifier is uncertain. For the generalized linear model, these samples are usually distributed around the current classification hyperplane. However, uncertain samples include samples near the current classification hyperplane, and samples far from the current classification hyperplane and the labeled samples. Traditional uncertainty sampling fails to describe the latter, and traditional methods are easily affected by outliers. In this paper, belief functions are used to describe the uncertainty that exists in various samples. Furthermore, we propose a sample selection strategy based on belief functions. Experimental results based on benchmark datasets show that the proposed approach outperforms several classical methods. Through this approach, higher classification accuracy can be achieved using the same number of new labeled samples.

Keywords active learning, uncertainty sampling, belief functions, generalized linear model

Citation Zhang S X, Han D Q, Yang Y. Active learning based on belief functions. Sci China Inf Sci, 2020, 63(11): 210205, https://doi.org/10.1007/s11432-020-3082-9

1 Introduction

Active learning (sometimes called query learning) is a subfield of machine learning. It selects a few critical unlabeled samples for manual labeling and then adds new labeled samples to the set of labeled samples for retraining the classifier. This process is conducted until the classifier's accuracy is achieved or the cost of labeling is exhausted. The whole process is shown in Figure 1.

Active learning has been widely used in applications such as medical image analysis [1], image restoration [2,3], and test classification [3,4]. The sample selection strategy plays a critical role in active learning (i.e., what kinds of samples are considered as critical samples or high-value samples). In general, sample selection strategies include information, representativeness, diversity and multicriteria-based strategies. The information-based strategies are used to select the samples for which the current classifier is uncertain. The use of different methods for measuring uncertainty leads to different samples being chosen. Lewis et al. [5] assumed that the sample with the least confidence was the most uncertain. However, Scheffer [6] claimed that the least confidence criterion only considered the information about the most probable label, discarding other labels' information. To solve this problem, Scheffer proposed margin sampling [6]. Considering all labels' information, the sample with maximum entropy would be regarded as the most uncertain one. Sharma et al. [7] further analyzed two reasons for the uncertainty of samples and proposed a probabilistic evidence-based uncertainty sampling. This method queried the sample

[*] Corresponding author (email: deqhan@gmail.com)

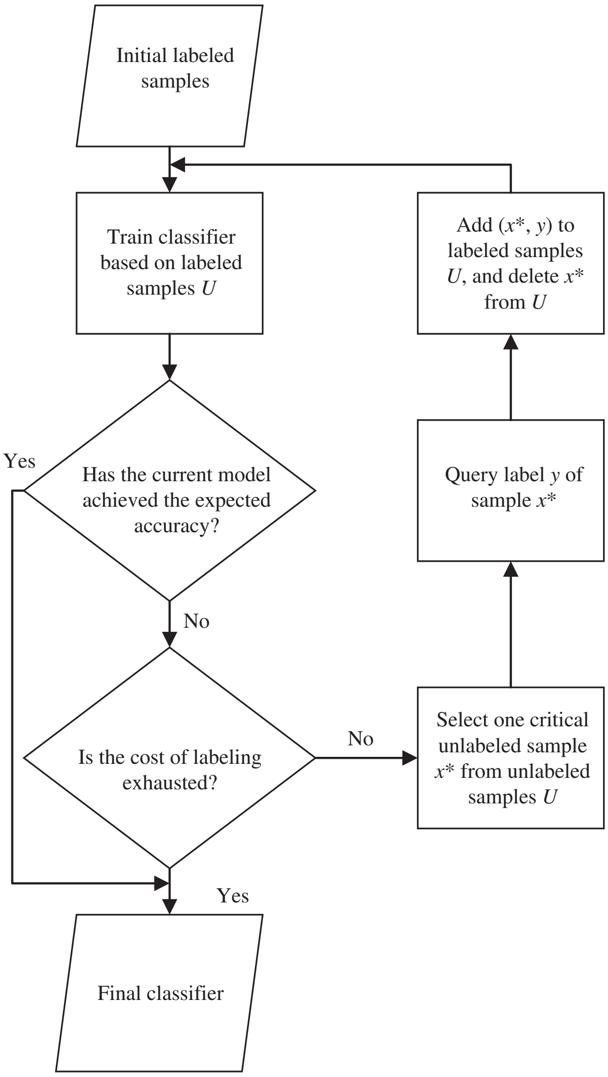

Initial labeled samples
Train classifier based on labeled samples U
Add (x*, y) to labeled samples U, and delete x* from U
Has the current model achieved the expected accuracy?
Yes
No
Query label y of sample x*
Is the cost of labeling exhausted?
No
Select one critical unlabeled sample x* from unlabeled samples U
Yes
Final classifier

Figure 1 The process of active learning.

with the most conflicting evidence from the most uncertain k samples. All the four methods mentioned above obtain uncertainty of unlabeled samples by calculating the output probability. When the output probability cannot be obtained directly, the distance between the sample and the current classification hyperplane is usually used to measure the uncertainty of the sample. In other words, a smaller distance leads to a larger uncertainty of the sample. Applying the distance uncertainty, Li et al. [8] proposed active learning based on multi-label support vector machine (SVM). Furthermore, to reduce the variance of the model, Zhang et al. [9] proposed to minimize the Fisher information ratio between $I_U(w)$ and $I_x(w)$, where $I_U(w)$ is the Fisher information matrix over the unlabeled pool U and $I_x(w)$ is the Fisher information matrix over x. For maximum model change, a sample that can lead to a great change of model is considered as critical sample. Cai et al. [10] used the gradient of the loss function to approximate the model change and derived algorithms for both SVM and logistic regression. Those sample selection strategies could be regarded as a variation of uncertainty sampling, which uses the norm of the sample, or the sample variance to weight the probability of the sample. Most of the samples obtained through the traditional information-based methods are distributed around the current classification hyperplane. Although the traditional information-based methods perform well in most cases, they still suffer from a few shortcomings.

(1) The problem of outliers: the outliers are the samples that are far from other samples. If outliers are selected, the performance of the classifier might be affected seriously. To reduce this influence, some

scholars further take account of the representativeness of the samples, considering the uncertainty of samples. Zhu et al. [11] used k-nearest neighbor (KNN) to select the representative samples. Hu et al. [12] selected a representative sample by hierarchical graph-theoretic clustering. Although the above algorithms have been proposed to combine informative and representative criteria, Huang et al. [13] thought these methods were ad hoc. From a min-max perspective, he selected samples characterized by both representativeness and informativeness.

(2) They cannot describe the uncertainty of samples from unknown regions well: these samples are far from the current classification hyperplane and the labeled samples, but are near to (or surrounded by) the unlabeled samples. The traditional information-based methods consider the uncertainty of these samples to be relatively small. Recently, with the rise of deep learning, there have been many methods based on neural networks that attempt to handle the selection of such samples. Zhu et al. [14] proposed a generative adversarial active learning (GAAL) method, which used deep convolution to generate an adversarial network to select critical samples. Sinha et al. [15] proposed to learn a latent space using a variational auto-encoder (VAE) and an adversarial network to select critical samples. The samples selected by deep models are usually those far away from the labeled ones. However, these methods have a huge computational complexity.

Traditional uncertainty sampling focuses on uncertain samples that are distributed around the classification hyperplane and ignores those from an unknown region. Besides, it might be affected by outliers. As a commonly used tool of uncertainty modeling and reasoning, the theory of belief functions can model the above two types of uncertain samples well and reduce the influence of outliers. Therefore, we use belief functions to model the uncertainty of samples and propose a sample selection strategy based on belief functions for active learning. Extensive experimental results show that the newly proposed approach is an effective improvement of the traditional uncertainty sampling.

2 Preliminaries

2.1 Sample selection strategy

Settles et al. [6] summarized the following existing sample selection strategies: uncertainty sampling, error reduction, variance reduction, minimization loss increase, maximum model change, adaptive method and among others. The purpose of this work is to explore the characteristics and improve the uncertainty criterion for active learning, hence minimizing loss increase and maximum error reduction. The adaptive methods based on multi-criteria are beyond the scope of this paper. Maximum model change and variance reduction can be regarded as variants of uncertainty sampling; thus, they are within the scope of discussion. The following are some detailed introductions to the comparison algorithm involved in this paper.

(1) The uncertainty sampling is the simplest and the most commonly used sample selection strategy for active learning. It can be used for probabilistic learning models directly. There are three main sample selection strategies based on uncertainty criterion in active learning.

(1.1) Least confidence (LC) selects the sample with the least confidence, which is the sample with the largest 0/1 loss:

$$x_{\mathrm{LC}} = \operatorname*{argmin}_{x_i}(1 - p(\hat{y}|x_i)), \tag{1}$$

where $\hat{y}$ is the label with the maximum posterior probability. LC only uses $p(\hat{y}|x_i)$ and loses the information on the remaining label distribution $p(y\backslash\hat{y}|x_i)$.

(1.2) Margin sampling (MS) considers the largest and the second largest posterior probabilities of the sample:

$$x_{\mathrm{MS}} = \operatorname*{argmin}_{x_i}(p(\hat{y}_1|x_i) - p(\hat{y}_2|x_i)), \tag{2}$$

where $\hat{y}_1$ represents the largest posterior probability label and $\hat{y}_2$ indicates the second largest posterior probability. Samples with small margin are more ambiguous. However, for multi-classification problem (the number of classes is greater than 2), MS still ignores much information for the remaining classes [6].

(1.3) A broader strategy is to use the probability entropy of unlabeled samples, which takes advantage of all the prediction information of the unlabeled sample:

$$x_{\text{Entropy}} = \underset{x_i}{\operatorname{argmin}} \sum_k p(y_k|x_i) \log(p(y_k|x_i)), \tag{3}$$

where $p(y_k|x_i)$ indicates the probability of the sample belonging to k.

The above method gets similar results in binary classification. They all simply query the samples whose posterior probability of being positive is nearest to 0.5. This paper selects the maximum entropy strategy as a representative of uncertainty sampling.

(2) The idea of variance reduction (VR) [16] selects the sample that can minimize the average variance of the estimated of model parameters. Zhang et al. [9] proposed using the Fisher information matrix over the unlabeled pool as the variance of the current model. Zhang's method can be computed as follows:

$$I_U = \frac{1}{|U|} \sum_{x_i \in U} p_1(x_i)(1 - p_1(x_i)) x_i x_i^{\mathrm{T}} + \lambda I_d, \tag{4}$$

where p_1 represents samples' probability belonging to the positive class and $|U|$ represents the number of unlabeled samples. Fisher information matrix over the unlabeled sample x_i can be computed as follows:

$$I_{x_i} = p_1(x_i)(1 - p_1(x_i)) x_i x_i^{\mathrm{T}} + \lambda I_d. \tag{5}$$

He minimized the Fisher information ratio between I_U and I_{x_i} to achieve VR:

$$x_{\text{VR}} = \underset{x_i \in U}{\operatorname{argmin}} \operatorname{tr}\left(I_{x_i}^{-1} I_U\right). \tag{6}$$

The more uncertain the sample, the greater the product of sample's probability. The greater the product of sample's probability is, the more Fisher information the sample has. Therefore, VR can be regarded as the variant of uncertainty criterion.

(3) Maximum model change (MMC) [17] selects the sample that can lead to great change of the current classifier. Settles used the expected gradient length of the objective function as the measurement of model change. Cai et al. [17] proposed that MMC based on logistics regression (LR) can be equivalent to the following:

$$x_{\text{MMC}} = \underset{x_i}{\operatorname{argmax}}\, 2p_1(x_i)(1 - p_1(x_i)) \|x_i\|. \tag{7}$$

Hence, MMC can be seen as a criterion using the norm to weight the uncertainty. This criterion based on SVM can be equivalent to the following:

$$x_{\text{MMC}} = \underset{x_i}{\operatorname{argmax}} \|x_i\| \quad \text{s.t.} \ \left|w^{\mathrm{T}} x_i + b\right| < \gamma, \tag{8}$$

where $\|x_i\|$ is the norm of x_i. The above methods can be regarded as uncertainty sampling or variants of uncertainty sampling. They are all used as comparison methods in this paper.

(4) Diversity criterion [18] selects the sample that is as dissimilar from the labeled samples as possible. In fact, this sample comes from the unlabeled area (the region where labeled samples are not distributed). Hence it can be regarded as an uncertain sample in essence. Kee et al. [12] uses the minimum distance between a sample and a labeled sample to measure the diversity of samples, as shown in the following formula:

$$x_{\text{Diversity}} = \underset{x_i}{\operatorname{argmin}} \|x_i - x_j\|^2, \tag{9}$$

where $x_j \in L$, L represents the set of labeled samples.

(5) Random sampling selects the sample in the set of unlabeled sample randomly. As the number of labeled samples increases, the distribution of labeled samples becomes closer to the distribution of original data and the performance of classifier improves. Random sampling is the baseline to improve classifier [6], and thus it is also selected as comparison algorithm in this paper.

2.2 The theory of belief functions

The theory of belief functions, also known as Dempster-Shafer evidence theory (DST), is a tool of uncertainty modeling and reasoning [19–21]. We briefly introduce the theory of belief functions in this paper as follows. Suppose A is a subset of a finite discrete frame of discernment (FOD) Θ, and then there exist

$$\begin{cases} m(\emptyset) = 0, \\ \sum\limits_{A \subseteq \Theta} (m(A)) = 1, \end{cases} \tag{10}$$

where $m : 2^{\Theta} \rightarrow [0,1]$ is called a basic belief assignment (BBA) defined over the FOD Θ for the closed-world assumption. Elements in FOD are mutually exclusive and exhaustive. In the open-world assumption, $m(\emptyset) > 0$ is allowed, which represents a mass assignment given to the hypothesis that might not lie in FOD. To get a better decision, the Dempster's rule of combination can be used to fuse those evidences while multiple independent evidences are available. The Dempster's rule of combination is expressed as follows:

$$m(A) = \begin{cases} 0, \quad A = \emptyset, \\ \dfrac{\sum_{\cap A_j = A} \prod_{1 \leqslant i \leqslant n} m_i\left(A_j\right)}{1-K}, \quad A \neq \emptyset, \end{cases} \tag{11}$$

where

$$K = \sum_{\cap A_j = \emptyset, 1 \leqslant i \leqslant n} m_i\left(A_j\right) \tag{12}$$

represents the total conflicting mass assignments. Suppose that m is a BBA defined on the FOD $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, and then m can be transformed to pignistic probability distribution as

$$\mathrm{BetP}(\theta) = \sum_{\theta \in A \subseteq \Theta} \frac{m(A)}{|A|}, \quad \forall A \subseteq \Theta, \tag{13}$$

where $|A|$ is the cardinality of A. Then, based on the pignistic probability, the ambiguity measure (AM) can be computed as

$$\mathrm{AM}(m) = -\sum_{\theta \subseteq \Theta} \mathrm{BetP}(\theta) \log_2 \mathrm{BetP}(\theta). \tag{14}$$

AM represents the degree of ambiguity incorporated in a BBA, which includes the discord and non-specificity. The discord represents the disagreement in choosing among different alternatives whereas the non-specificity represents that two or more choices are left unspecified.

3 The proposed method

3.1 Problem statement

We use a binary classification to explain the drawback of the traditional uncertainty sampling. As shown in Figure 2, the red circle represents the labeled positive samples, the blue circle represents the labeled negative samples, the purple straight line represents the current classification hyperplane based on the labeled samples, and the cross represents the unlabeled samples. Based on the output probability of the unlabeled samples, the uncertainty of the samples (probability entropy; see Table 1) can be calculated. When the classifier is logistic regression, the probability is calculated as follows:

$$p(y=1|x_i) = \frac{1}{1+\exp(-w^{\mathrm{T}} x_i - b)}, \quad p(y=0|x_i) = \frac{\exp(-w^{\mathrm{T}} x_i - b)}{1+\exp(-w^{\mathrm{T}} x_i - b)}. \tag{15}$$

When the classifier is SVM, the sigmoid function is used, which mapped score $w^{\mathrm{T}} x_i + b$ to the probability. The probability is calculated as follows:

$$p(y=1|x_i) = \frac{1}{1+\exp(A(w^{\mathrm{T}} x_i + b) + B)}, \quad p(y=-1|x_i) = \frac{\exp(A(w^{\mathrm{T}} x_i + b) + B)}{1+\exp(A(w^{\mathrm{T}} x_i + b) + B)}, \tag{16}$$

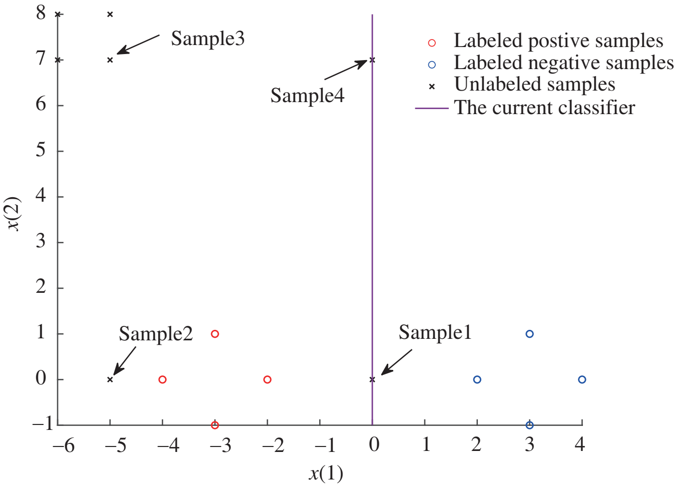


Figure 2 (Color online) The problem of traditional uncertainty sampling.

Table 1 Uncertainty based on probability entropy

	p_1	p_2	Probability entropy
Sample1	0.5	0.5	**0.6931**
Sample2	0.9933	0.0067	0.0402
Sample3	0.9933	0.0067	0.0402
Sample4	0.5	0.5	**0.6931**

where the parameters A and B correspond to the fields' scale and intercept of score transform and they are obtained by maximum likelihood. For logistic regression and SVM, $p(y=1|x_i)$ is written as $p_1(x_i)$ for simplicity, and $p(y=0|x_i)$ and $p(y=-1|x_i)$ are written as $p_2(x_i)$.

(1) The traditional uncertainty sampling will select Sample1 or Sample4 for active learning. As Sample1 and Sample4 are close to the current classification hyperplane, they are considered as uncertain samples. However, it can be clearly observed that Sample4 is an outlier. Therefore, the traditional method might select an outlier.

(2) The probability uncertainty of Sample2 is the same as that of Sample3, because they have the same distance to the classification hyperplane. Sample2 is close to the positive labeled samples; however, Sample3 is far away from the labeled samples and located in an unknown region. It is counter-intuitive that the probability uncertainty of Sample2 is the same as that of Sample3. The uncertainty of Sample3 should be greater than that of Sample2. Therefore, the traditional probability uncertainty cannot describe the uncertainty of Sample3 well.

3.2 BBA generation

Let us describe the uncertainty of the sample in terms of belief functions. For simplicity, the belief is only assigned to the singleton, the empty set, and the the total set. We discount the output probability of the sample and assign it to the singleton. Intuitively, if an unlabeled sample is similar to one of the labeled samples, then the output probability of the unlabeled sample is more credible and it is given a small discount to assign the singletons. Because the unlabeled sample is distributed in an unknown region and it is uncertain, the discount should be big so that more belief is assigned to the empty set or the total set. If the unlabeled sample is located in a dense region (in other words, it is near to other samples), it is less likely to belong to an outlier and $m_\emptyset$ should be small; otherwise, $m_\emptyset$ should be big. Then the remaining belief is assigned to the total set. Take a binary classification problem as an example. Based on this idea, the BBA can be generated by the following formula:

$$m_{x_i}(\theta_1) = f(x_i)p_1(x_i), \quad m_{x_i}(\theta_2) = f(x_i)p_2(x_i), \tag{17}$$

Table 2 Symbolic representation involved in the algorithm

Symbolic	Description	Symbolic	Description
L	All labeled samples	U	All unlabeled samples
x^*	Selected samples	$\mathrm{AM}(x_i^u)$	The ambiguity measure of x_i^u
$\phi_c(\cdot\|L)$	Classifier trained on labeled samples	$p(x_i^u)$	The output probability of x_i^u
$\|U\|$	The number of unlabeled samples	x_i^u	An unlabeled sample
$\mathrm{Un}(x_i^u)$	The uncertainty of x_i^u	$m_{x_i^u}(\theta)$	The belief function of x_i^u
η	Threshold about $m_\emptyset$		

where $f(x_i) = \mathrm{e}^{-\alpha d_l(x_i)}$ is the discount factor; $d_l(x_i) = \min_{x_j \in L} d(x_i, x_j)$ represents the Euclidean distance between x_i and the labeled sample that is the nearest to x_i; α is a parameter, whose value should be greater than or equal to 0, with the default value being 1. $p_1(x_i)$ is the probability that the sample belongs to the positive class and $p_2(x_i)$ is the probability that the sample belongs to the negative class.

$$m_{x_i}(\emptyset) = g(x_i)\left(1 - m_{x_i}(\theta_1) - m_{x_i}(\theta_2)\right), \tag{18}$$

where $g(x_i) = \mathrm{e}^{-\beta d_o(x_i)}$ is the density factor; $d_o(x_i) = \min_{x_k \in [L,U]\backslash x_i} d(x_i, x_k)$ represents the Euclidean distance between x_i and the sample (including the labeled samples and unlabeled samples) that is the nearest to x_i; β is a parameter, whose value should be greater than or equal to 0, with the default value being 1.

$$m_{x_i}(\Theta) = 1 - m_{x_i}(\theta_1) - m_{x_i}(\theta_2) - m_{x_i}(\emptyset). \tag{19}$$

Note that AM is calculated based on a closed-world assumption in Subsection 2.2. We make the following corrections to the mass calculated by the above method and calculate AM:

$$\begin{cases} m'_{x_i}(A) = \frac{m_{x_i}(A)}{1-m_{x_i}(A)}, & \text{if } A \in 2^\Theta \backslash \emptyset, \\ m'_{x_i}(\emptyset) = 0. \end{cases} \tag{20}$$

3.3 Uncertainty sampling based on belief functions (USBF)

Our sample selection strategy will be introduced in this subsection. Table 2 is the explanation of some symbols in Algorithm 1.

Algorithm 1 Traditional uncertainty sampling

```
Require: A set of labeled samples $L$, a set of unlabeled samples $U$.
while Termination condition not satisfied do
    Train a classifier $\phi_c(\cdot|L)$ based on labeled samples;
    for $i = 1 : |U|$ do
        Calculate the uncertainty of the sample, $\mathrm{Un}(x_i^u)$;
    end for
    $x^* = \mathrm{argmax}_i\, \mathrm{Un}(x_i^u)$;
    $U = U - x^*$;
    $L = L \cup (x^*, \mathrm{GetLabel}(x^*))$;
end while
```

For the traditional uncertainty sampling: first, it calculates the uncertainty of unlabeled samples (include the probability entropy and the other variant of the probability uncertainty; see Subsection 2.2 for detail); then, it selects the most uncertain sample to label and repeat this process. As mentioned above, traditional methods only focus on the samples that are distributed around the classification hyperplane as the uncertain samples, and they ignore that the one far from the classification hyperplane could be an uncertain sample. Additionally, the traditional methods might select an outlier. A method based on belief functions is proposed to address the issue of traditional strategy. First, the output probability of the unlabeled samples is still calculated; then, the belief functions are computed by the method in Subsection 2.2; next, the samples with larger $m_\emptyset$ are removed, since these samples are far away from others and are suspected to be the outliers; finally, we calculate the AM of the remaining samples and select the

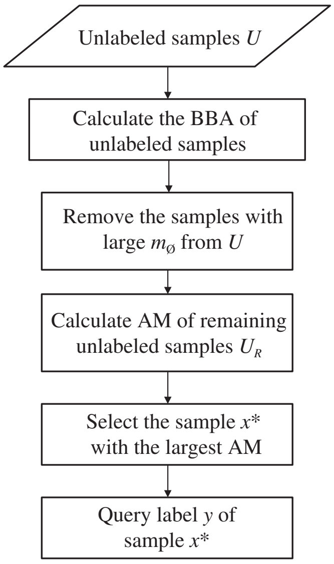

Unlabeled samples U
Calculate the BBA of unlabeled samples
Remove the samples with large $m_ø$ from U
Calculate AM of remaining unlabeled samples U_R
Select the sample x* with the largest AM
Query label y of sample x*

Figure 3 Uncertainty sampling based on belief functions.

Table 3 Uncertainty based on belief functions

	$m(\emptyset)$	$m(\theta_1)$	$m(\theta_2)$	$m(\Theta)$	AM
Sample1	0.1146	0.1839	0.1839	0.5175	**0.6931**
Sample2	0.0374	0.6025	0.0041	0.3560	0.4850
Sample3	0.0911	0.0420	2.833×E−4	0.8665	**0.6921**
Sample4	**0.3739**	0.0175	0.0175	0.5853	–

sample with the largest AM to label. USBF selects the samples with maximum AM for labeling; hence it is also called as MaxAM. The process of sample selection is shown in Figure 3. The whole algorithm process is shown in Algorithm 2.

Algorithm 2 USBF

```
Require: A set of labeled samples $L$, a set of unlabeled samples $U$.
while Termination condition not satisfied do
    Train a classifier $\phi_c(\cdot|L)$ based on labeled samples;
    $U_R = \emptyset$;
    for $i = 1 : |U|$ do
        $p(x_i^u) = \phi_c(x_i^u|L)$;
        Calculate the belief functions of $x_i^u$,$m_{x_i^u}(\theta)$;
        if $(m_{x_i^u}(\emptyset) < \eta)$ then
            $U_R = U_R + x_i^u$;
        end if
    end for
    for $i = 1 : |U_R|$ do
        Calculate the AM$(x_i^u)$;
    end for
    $x^* = \underset{i}{\operatorname{argmax}}\,\mathrm{AM}(x_i^u)$;
    $U = U - x^*$;
    $L = L \cup (x^*, \mathrm{GetLabel}(x^*))$;
end while
```

Let us calculate the BBA of the unlabeled samples in Figure 2 based on the above method. The calculation results are presented in Table 3. It is obvious that $m_\emptyset$ of Sample4 is relatively large; hence Sample4 is very likely to be an outlier. We remove Sample4 and calculate AM for other samples. The uncertainty of Sample1 and Sample3 will be larger, and thus they are considered as uncertain samples.

Table 4 Details of datasets

Data	Category	Num	Proportion	Features
Heart	2	270	150/120	13
Breast	2	683	444/239	9
LetterDP	2	1608	805/803	16
LetterIJ	2	1502	755/747	16
LetterVY	2	1550	764/786	16
LetterEF	2	1543	768/755	16
7VS9	2	14251	7293/6958	784(PCA10)
Toy1	2	1000	500/500	2
Toy2	2	400	200/200	2

4 Experimental analysis

To verify the effectiveness of our algorithm, we compare our method based on USBF with the classical uncertainty sampling, maximum model change, maximum variance reduction, the method based on diversity, and random sampling in 7 public datasets and 2 toy datasets. The abbreviations and descriptions of these methods are as follows.

(1) Random: random sampling, selecting a sample in the unlabeled set randomly.

(2) MaxEntropy: maximum probability entropy, selecting a sample with maximum entropy (see for formula (3)).

(3) MaxAM: maximum ambiguity measure, selecting a sample with maximum AM (see formula (14)).

(4) MMC: maximum model change, selecting a sample that can lead to a great change of current model once labeled (see formula (8)).

(5) VR: variance reduction, selecting a sample that can reduce the average variance of the estimates of model parameters (see formula (6)).

(6) Diversity: selecting a sample which is not similar to the labeled samples (see formula (9)).

4.1 Datasets

To compare the difference between different sample selection strategies for active learning, 9 datasets were used to verify the validity of the algorithm. Some basic information and pretreatment of the datasets are shown in Table 4. For the Letter dataset, we converted it into a binary classification problem, including four groups, namely LetterIJ, LetterVY, LetterEF and LetterDP. For the MINIST dataset, it was also converted into a binary-category problem, including one group, namely 7VS9; principal component analysis (PCA) was used to reduce the feature dimension from original 784 to 10. One of the toy datasets, shown as Figure 4, is a two-dimensional normal distribution with mean of -5, 5 and a variance of $\sqrt{5}$. The other is shown in Figure 5. In this dataset [13], if the initial labeled sample is not selected properly, the traditional uncertainty sampling will fall into local optimum.

4.2 Datasets division and evaluation criteria

Original dataset was divided into a labeled set, an unlabeled set, and a test set randomly. For the initial labeled set, the number of samples is at least k, where k is the number of classes contained in the dataset, and there is at least one sample for each class. 30 Monte-Carlo experiments were conducted on each dataset, and then the final result was obtained by taking the average result of 30 experiments. The number of selecting samples was 30, and one sample was selected at one iteration process. We draw the learning curve and use the area under the learning curve (ALC) as performance measurement. One-factor analysis of variance was carried out at the level of 95% significance.

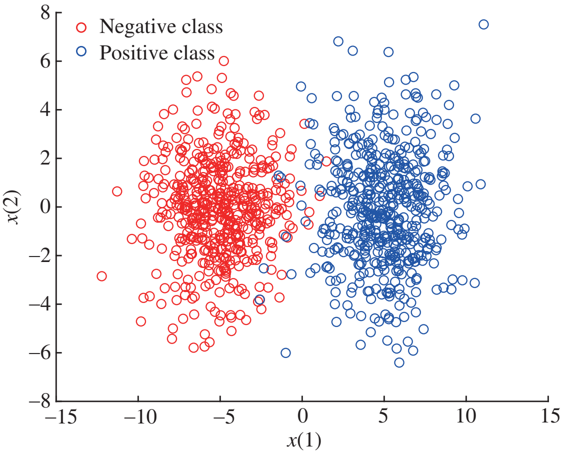


Figure 4 (Color online) Distribution of artificial dataset Toy1.

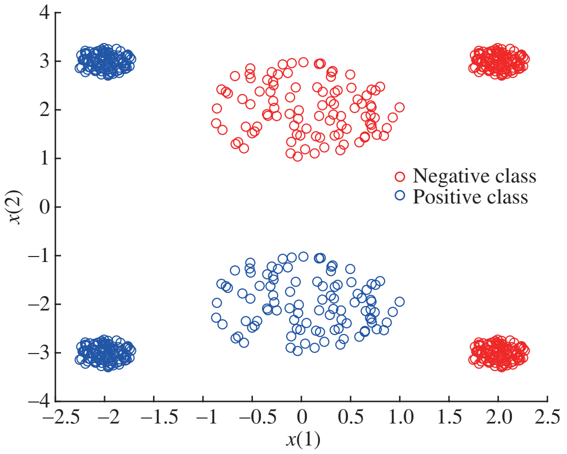


Figure 5 (Color online) Distribution of artificial dataset Toy2.

Table 5 Performance comparison of active learning algorithms in terms of ALC (the classifier is LR)[a)]

	Random	MaxEntropy	MaxAM	MMC	VR	Diversity	Pr
Heart	20.7303	21.3208	**22.5876**	21.3792	20.9729	22.1399	<0.001
Breast	23.8481	24.1926	**24.3251**	23.6402	24.1977	23.6796	<0.001
DvsP	27.3869	28.5772	**28.6395**	28.6394	27.6903	28.2065	<0.001
IvsJ	25.2297	26.1200	**26.2846**	25.5102	25.6006	25.4774	<0.001
VvsY	24.8488	25.5989	**25.7868**	25.5066	25.1008	25.0405	0.281
EvsF	27.1054	28.1053	**28.1780**	27.4638	27.2350	27.6923	<0.001
7VS9	26.7563	26.9520	**26.9590**	26.7247	26.8880	26.8400	<0.001
Toy1	29.1916	29.3444	29.3830	**29.3839**	29.3553	29.3433	<0.001
Toy2	29.1606	27.1131	**29.7385**	27.2998	27.2559	29.1533	<0.001
Win	0	0	**8**	1	0	0	–
Rank	5.20	3.11	**1.11**	3.67	3.89	4	–

a) The best algorithm on each dataset has been highlighted in bold face; Win is the number of datasets on which an algorithm achieves the best or comparable; Rank shows the average rank within the compared methods.

4.3 Experimental environment

We performed experiments on a computer equipped with Intel Core i7 3.6 GHz CPU, 8 GB DDR III memory and MATLAB 2018b software of Microsoft Windows 10 OS. In Subsection 3.1, we notice that generalized linear models often use the distance from the sample to the classification hyperplane as the sample uncertainty directly or indirectly. This measurement method has certain defects; therefore, USBF is proposed to improve it. LR and SVM are representative of generalized linear models, so they were chosen as classifiers. We use the LR contained in the liblinear package and set the regularization parameter λ to 0.01 [22]. For SVM, the probability formula is shown in formula (16).

4.4 Result analysis

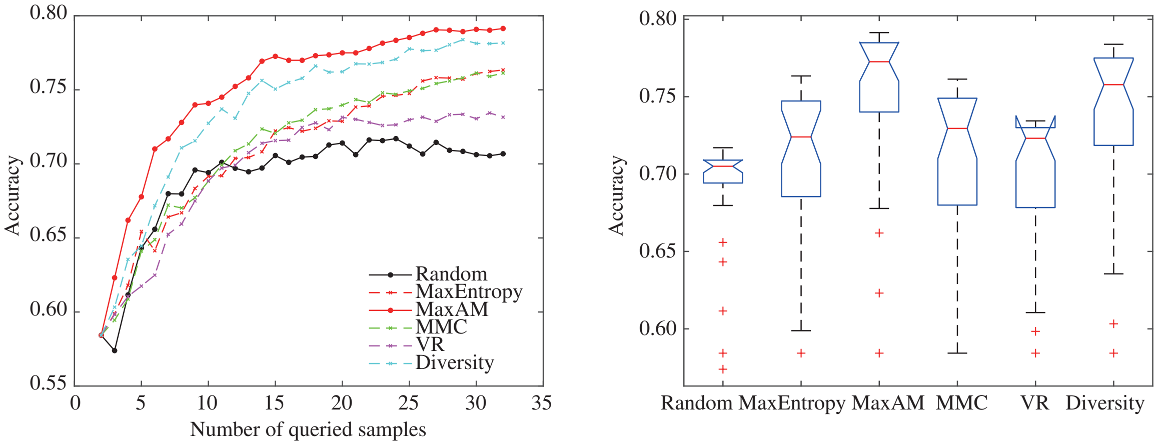
Tables 5 and 6 show the performance of various sample selection strategies under different classifiers. Figures 6–14 show the performance of active learning on each dataset based on LR and Figure 15 shows the average learning curve of the above datasets. Figure 16 shows the average learning curve of the above datasets when the classifier is SVM. Owing to space limitations, only their average results are listed. The average run time of each method is shown in Table 7.

For LR, we notice that Random, VR, and Diversity do not achieve the best performance on all datasets. Their rankings are 5.20, 3.89 and 4.00, respectively. MaxEntropy ranks second overall. MaxAM tends to yield decent performance in most cases, with an average rank of 1.11 and achieves the best improvement

Table 6 Performance comparison of active learning algorithms in terms of ALC (the classifier is SVM)$^{a)}$

	Random	MaxEntropy	MaxAM	MMC	VR	Diversity	Pr
Heart	19.9130	19.8200	**20.4494**	19.6234	19.6234	19.9469	0.6839
Breast	23.1817	23.5343	**23.8023**	23.5646	23.4649	23.6947	<0.001
DvsP	28.0414	28.4266	**28.7438**	28.3654	28.0947	28.2963	<0.001
IvsJ	26.7640	25.5787	**25.8420**	25.1469	24.8652	25.3375	<0.001
VvsY	24.3555	25.2773	**25.6406**	24.8179	24.0420	24.1685	<0.001
EvsF	27.4661	28.1299	**28.2953**	27.7841	27.5363	27.4708	<0.001
7VS9	26.5471	26.8171	**26.8912**	26.6995	26.7957	26.5487	<0.001
Toy1	29.2302	29.4354	**29.4426**	29.3466	29.4112	29.3556	<0.001
Toy2	26.8595	26.4866	**28.5427**	27.2929	26.2620	28.0407	<0.001
Win	0	0	**9**	0	0	0	–
Rank	5.38	2.5	**1**	3.88	4.63	3.75	–

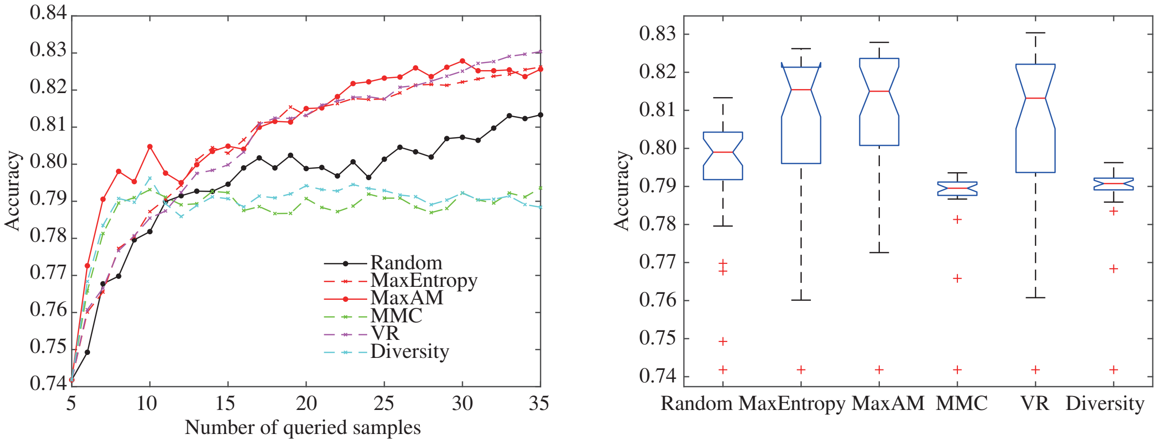
a) The best algorithm on each dataset has been highlighted in bold face; Win is the number of datasets on which an algorithm achieves the best or comparable; Rank shows the average rank within the compared methods.



(a)

(b)

Figure 6 (Color online) Active learning performance on Heart dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

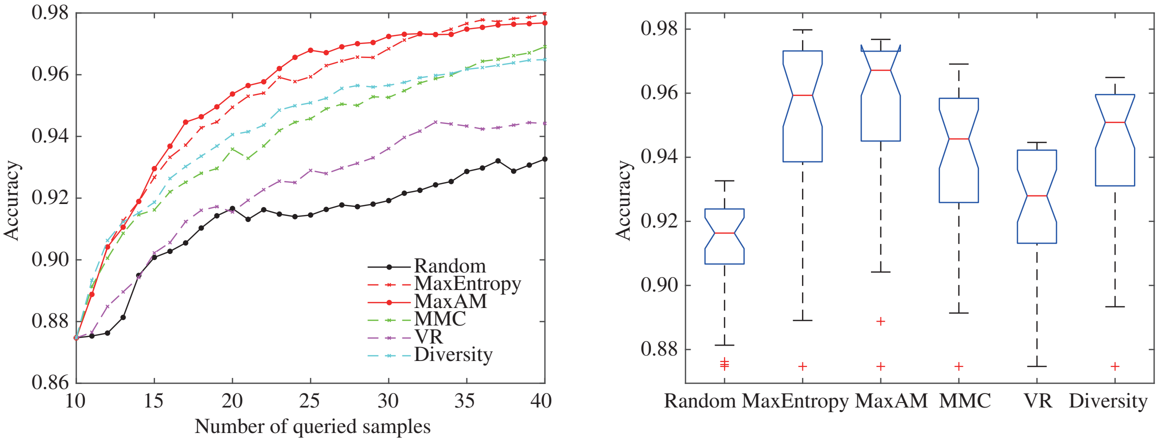


(a)

(b)

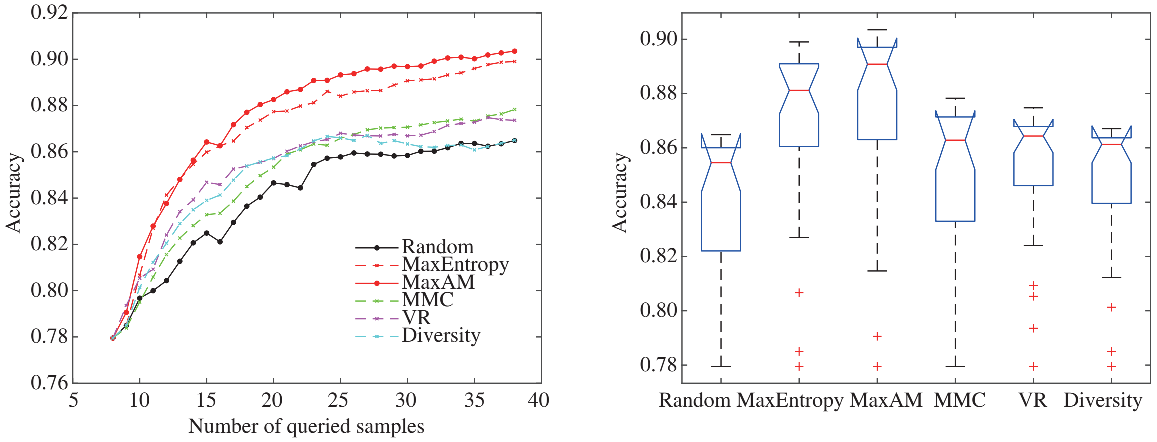
Figure 7 (Color online) Active learning performance on Breast dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

8 times. MMC achieves the best performance 1 times, whose average rank is 3.67. It has beaten MaxAM on Toy1, but the difference between them is not obvious. This phenomenon can be attributed to the



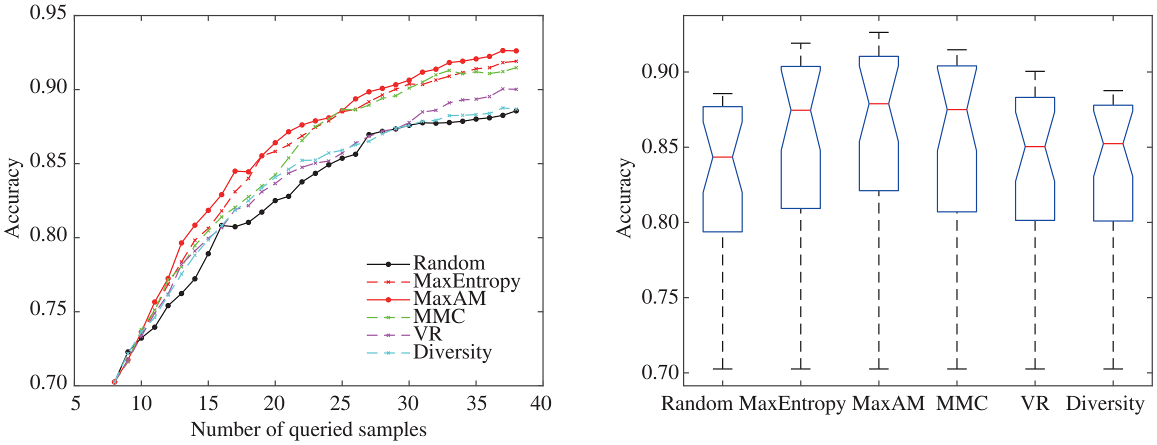
(a) (b)

Figure 8 (Color online) Active learning performance on LetterDP dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.



(a) (b)

Figure 9 (Color online) Active learning performance on LetterIJ dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.



(a) (b)

Figure 10 (Color online) Active learning performance on LetterVY dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

fact that this dataset is relatively simple and there is no problem of falling into a local optimum or an outiler. In contrast to Toy2, if the initial labeled samples are selected from the central area as shown

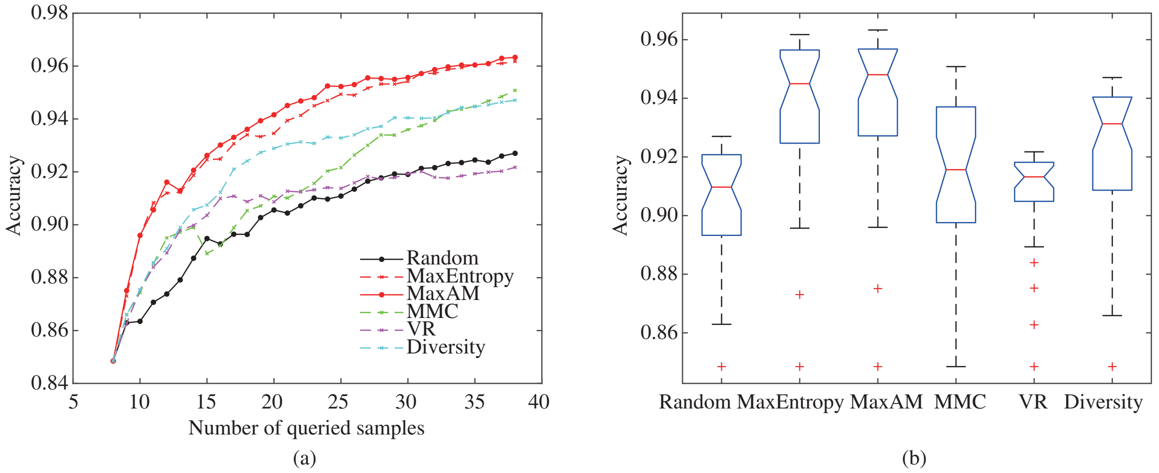


Figure 11 (Color online) Active learning performance on LetterEF dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

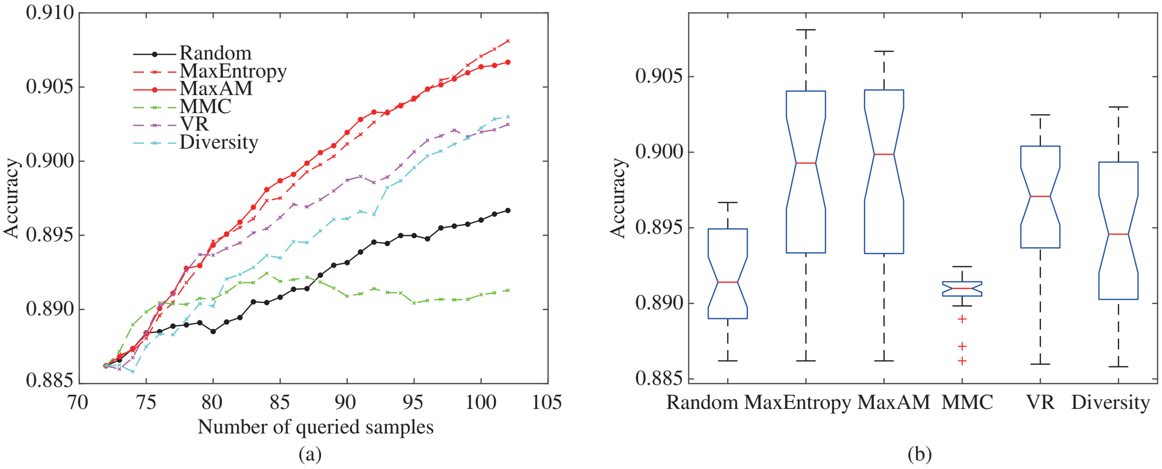


Figure 12 (Color online) Active learning performance on 7VS9 dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

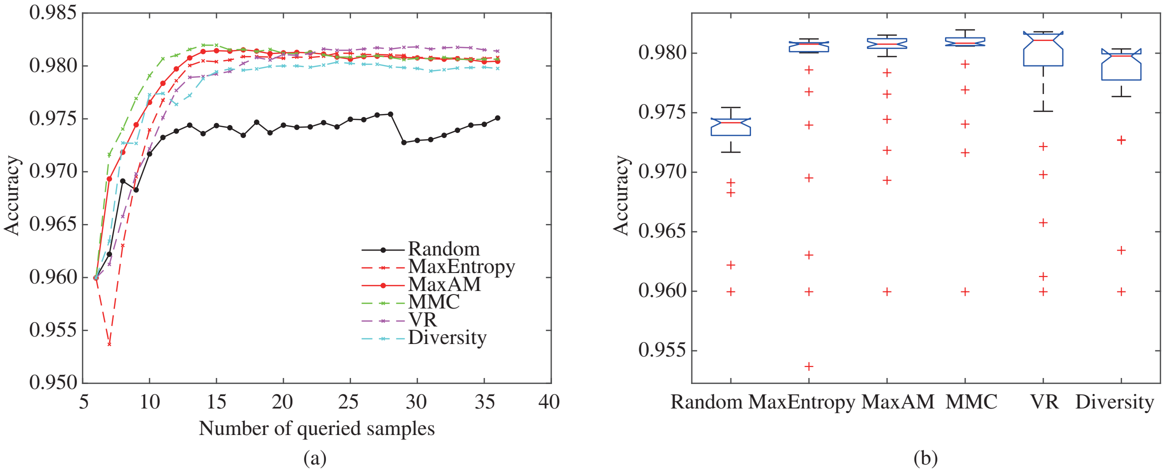
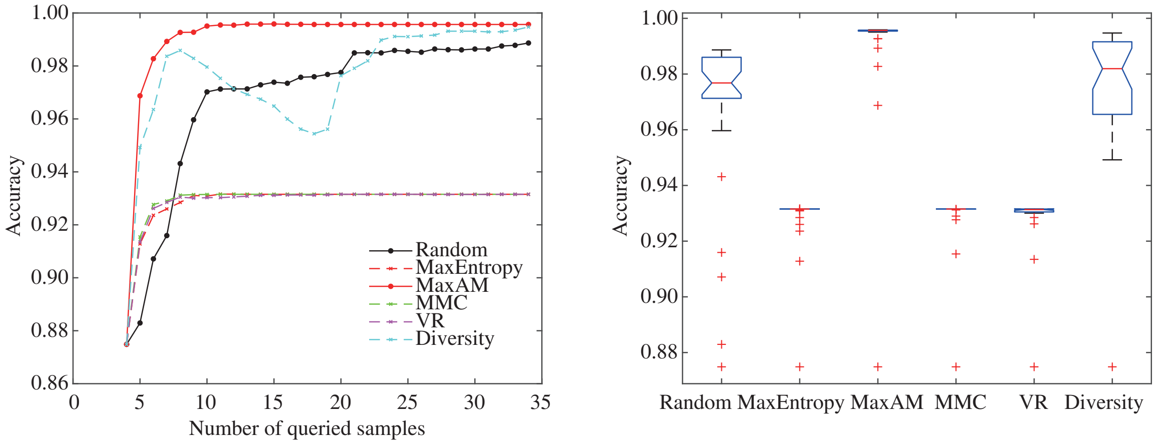


Figure 13 (Color online) Active learning performance on Toy1 dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

in Figure 5, the data from the area on the four corners will not be selected. Because they are far from the current classification hyperplane, traditional uncertainty measures will not consider them to



(a) (b)

Figure 14 (Color online) Active learning performance on Toy2 dataset based on LR. (a) Learning curve; (b) distributions of the classifier's accuracy.

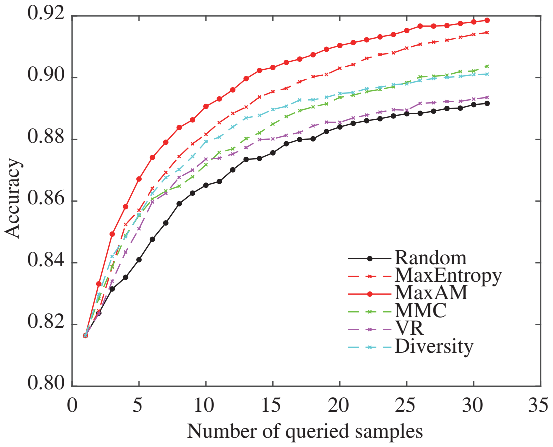


Figure 15 (Color online) Average ALC based on LR.

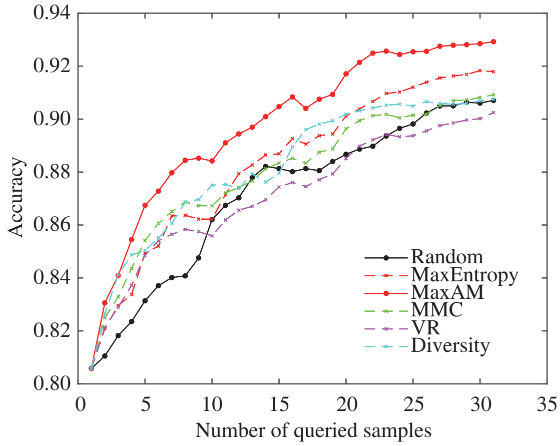


Figure 16 (Color online) Average ALC based on SVM.

Table 7 Average run time of active learning algorithms

	Random	MaxEntropy	MaxAM	MMC	VR	Diversity
LR	0.023	0.026	**0.645**	0.052	0.263	0.058
SVM	0.024	0.828	**1.687**	0.923	1.122	0.056

be uncertainty samples. In fact, for high-dimensional datasets, the dataset often exhibits multi-peak characteristics (samples of the same category are distributed in multiple regions). In this case, many unknown regions will not be queried since they are far from the classification hyperplane, which leads to local optimization. Their performance is not even as effective as random sampling. MMC and VR will also encounter the same problem. Diversity method will not encounter this problem, but it might select an outlier. MaxAM based on the belief functions can reduce the influence of those problem; hence the behavior of MaxAM outperforms other methods. For SVM, the performance of each method is roughly the same as that based on LR. MaxAM achieves the best performance on all datasets.

Regarding training time, no matter which classifier is selected, Random and Diversity methods take the least time. It is because that the above methods are completely based on the data itself and the sample selection process does not need to train the classifier. MaxAM takes the most time. For calculating the AM of the sample, it not only requires the current classifier to participate in it, but also transforms the output probability of the sample into BBA. It leads to a lot of time overhead.

5 Conclusion

This paper proposes a new sample selection strategy for active learning: USBF. The belief functions are used to model the uncertainty of the sample. The samples near the classification hyperplane are considered as uncertain samples; those far from the classification hyperplane and labeled samples are also considered to be uncertain samples. Traditional uncertainty sampling methods fail to model the latter. The introduction of empty sets allows us to reduce the effect of outliers effectively. Based on LR and SVM, USBF can produce the best results using large number of datasets. We recommend using belief functions to measure the uncertainty of the sample. The sample selection criteria, e.g., information, representative, divergence and multi-criteria based strategies, can be employed for active learning. Combining representative and diversity criteria to improve the performance of current classifiers better is the subject of future work. Additionally, generating BBA has a large computational complexity when FOD is large. Under the premise of ensuring the accuracy of the algorithm, reducing the computational complexity is also a challenging problem.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61671370), Postdoctoral Science Foundation of China (Grant No. 2016M592790), and Postdoctoral Science Research Foundation of Shaanxi Province (Grant No. 2016BSHEDZZ46).

References

1 Zhou Z W, Shin J, Zhang L, et al. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: Proceedings of Computer Vision and Pattern Recognition, 2017. 4761–4772

2 Hoi S C H, Rong J, Zhu J K, et al. Semi-supervised SVM batch mode active learning for image retrieval. In: Proceedings of Computer Vision and Pattern Recognition, 2008. 1–7

3 Hoi S C H, Rong J, Lyu M R. Batch mode active learning with applications to text categorization and image retrieval. IEEE Trans Knowl Data Eng, 2009, 21: 1233–1248

4 Raghavan H, Madani O, Jones R. Active learning with feedback on features and instances. J Mach Learn Res, 2006, 7: 1655–1686

5 Lewis D D, Catlett J. Heterogenous uncertainty sampling for supervised learning. In: Proceedings of the 11th International Conferenceon on Machine Learning, 1994. 148–156

6 Settles B. Active Learning Literature Survey. Technical Report, Department of Computer Science, University of Wisconsin-Madison. 2010

7 Sharma M, Bilgic M. Evidence-based uncertainty sampling for active learning. Data Min Knowl Disc, 2017, 31: 164–202

8 Li X, Guo Y H. Active learning with multi-label SVM classification. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013. 1479–1485

9 Zhang T, Oles F. The value of unlabeled data for classification problems. In: Proceedings of the 17th International Conference on Machine Learning, 2000. 1191–1198

10 Cai W B, Zhang Y X, Zhang Y, et al. Active learning for classification with maximum model change. ACM Trans Inf Syst, 2017, 36: 1–28

11 Zhu J, Wang H, Yao T, et al. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008. 18–22

12 Kee S, del Castillo E, Runger G. Query-by-committee improvement with diversity and density in batch active learning. Inf Sci, 2018, 454–455: 401–418

13 Huang S J, Jin R, Zhou Z H. Active learning by querying informative and representative examples. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 1936–1949

14 Zhu J J, Bento J. Generative adversarial active learning. 2017. ArXiv: 1702.07956

15 Sinha S, Ebrahimi S, Darrell T. Variational adversarial active learning. 2019. ArXiv: 1904.00370

16 Yang Y, Loog M. A variance maximization criterion for active learning. Pattern Recogn, 2018, 78: 358–370

17 Cai W B, Zhang Y, Zhou S Y, et al. Active learning for support vector machines with maximum model change. Machine Learn Knowl Discov Databases, 2014, 9: 211–226

18 Zhu J, Wang H, Tsou B K, et al. Active learning with sampling by uncertainty and density for data annotations. IEEE Trans Audio Speech Lang Process, 2010, 18: 1323–1331

19 Masson M H, Denœux T. ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recogn, 2008, 41: 1384–1397

20 Han D Q, Liu W B, Dezert J, et al. A novel approach to pre-extracting support vectors based on the theory of belief functions. Knowledge-Based Syst, 2016, 110: 210–223

21 Smets P, Kennes R. The transferable belief model. Artif Intell, 1994, 66: 191–234

22 Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: a library for large linear classification. J Mach Learn Res, 2008, 9: 1871–1874