A BIC Based Initial Training Set Selection Algorithm for Active Learning and Its Application in Audio Detection

Yan LENG [1], Guang-hui QI [2], Xin-yan XU [3]

[1] Dept. of Physics and Electronics, Shandong Normal University, East Wenhua Road 88, 250014, Ji'nan, China
[2] Dept. of Mechanical Engineering, Shandong Jiaotong University, Jiaoxiao Road 5, 250023, Ji'nan, China
[3] Dept. of Computer Science and Technology, Shandong College of Electronic Technology, Wenhua Road 678, 250200, Zhangqiu, Ji'nan, China

lyansdu@163.com, guanghui-qi@163.com, xuxinyan@sohu.com

Abstract. *To construct a classification system or a detection system, large amounts of labeled samples are needed. However, manual labeling is dull and time consuming, so researchers have proposed the active learning technology. The initial training set selection is the first step of an active learning process, but currently there have been few studies on it. Most active learning algorithms adopt random sampling or algorithms like sampling by clustering (SBC) to select the initial training samples. But these two kinds of method would lose their effectiveness in detecting events of small probability because sometimes they could not select or select too few samples of the small probability events. To solve this problem, this paper proposes a BIC based initial training set selection algorithm. The BIC based algorithm performs clustering on the whole training set first, then uses BIC to judge the status of clusters. Finally, it adopts different selection strategies for clusters of different status. Experimental results on two real data sets show that, compared to random sampling and SBC, the proposed BIC based initial training set selection algorithm can efficiently solve the detection problem of small probability events. In the mean time, it has obvious advantages in detecting events of non-small probability.*

Keywords

Initial training set selection, active learning, BIC, subspace sample selection, audio detection.

1. Introduction

In many fields, constructing a classification system or a detection system needs large amounts of labeled training samples. While manual labeling would cost people lots of time and energy, which makes it very expensive to obtain labeled samples. For example, in the audio detection field, in order to detect different audio events in the continuous audio stream, we need to label the audio events contained in the audio stream one by one. As another example, in the web search field, it is unrealistic to let the user hand-label a thousand training pages as interesting or not in order to find the web pages that the user is interested in. When detecting events of small probability, the expensive labeling problem would be more serious. For example, in the audio detection field, in an audio stream, the proportion of a small probability event is too small. Also the small probability events would be scattered in time domain. Then we would have obtained very few samples of small probability events after labeling a very long audio stream. To solve the above problem, researchers have proposed the active learning (AL) technology, and have done significant research on it [1-8]. AL is to query the samples that are most informative for training. So compared to passive learning, it can reduce manual labeling workload. There are mainly three issues in AL. First, a small initial training set should be selected to start the AL process. Second, a sampling strategy is required to choose the informative samples for manual labeling. Third, a stopping criterion should be established to determine when to stop the AL process. Lots of researches have been done on the second issue, while the first issue has received little consideration. In this paper, we focus on the first issue, that is, the initial training set selection, and propose a BIC based initial training set selection algorithm for AL. BIC is the abbreviation of Bayesian Information Criterion which will be introduced in Section 4.1.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 states the problems existing in current initial training set selection algorithms. Section 4 presents our BIC based initial training set selection algorithm. Section 5 shows experimental results and analysis. Section 6 gives conclusions.

2. Related Work

AL can be mainly divided into two categories: uncertainty sampling [9] and committee-based sampling [10], of which uncertainty sampling is the mostly used method. For uncertainty sampling, an active learner can be modeled as (C, Q, E, L, U). C is a classifier that is trained by the labeled training set L. Q is the query function that is used

to select the informative samples from the unlabeled set U. E is the expert that can assign true labels to the selected samples. The procedure of uncertainty sampling AL is:

① A few initial training samples are selected and are hand-labeled by the expert to generate the initial training set L.

② L is used to train an initial classifier C to seed the AL process.

③ The query function Q is used to select informative samples from U.

④ The informative samples are hand-labeled by the expert E and are put into L.

⑤ The updated training set L is then used to retrain the classifier C.

⑥ Go to ③. The iteration continues until achieving a predefined stopping criterion.

Currently, most researchers focus on designing the query function Q, while ignore the study of the initial training set selection. Actually, initial training set selection is an important part of AL. Experiments in literature [11] show that the initial training set would have a great impact on the convergence rate of AL.

Most current AL algorithms adopt random sampling to generate the initial training set. Hu et al. [12] reviewed 206 papers about AL from conferences including NIPS, ICCV, ICML etc. and from journals including Machine Learning, Pattern Recognition etc., and finally concluded that over 94% researchers had adopted a randomly selected initial training set or had failed to specify their initial training set selection methods. The initial training set is generally generated by random sampling based on the assumption that random sampling will be likely to build the initial training set with the same prior distribution as that of the whole training set. However, Zhu et al. [13] point out that the above assumption seldom holds in reality since the size of the initial training set is much smaller than that of the whole training set. In this case, they emphasize the representativeness of the initial training set, and propose a sampling by clustering (SBC) method. SBC first uses k-means to cluster all unlabeled samples into a predefined number of clusters, and then for each cluster, the sample closest to the cluster centroid is taken as the representative one, and is selected to augment the initial training set. Other initial training set selection algorithms have a similar idea as SBC. For example, Kang et al. [14] take the method just like SBC to select the initial training samples in each cluster. Moreover, they also put the centroids themselves into the initial training set. Nguyen et al. [5], Yuan et al. [15], Cebron et al. [16], and Hu et al. [12] also take the methods similar to SBC to generate the initial training set, but use different clustering algorithms. Nguyen et al. use k-medoid. Yuan et al. and Cebron et al. use fuzzy c-means, and Hu et al. use FFT, AHC and APC. In summary, the current initial training set selection methods can be mainly divided into two categories: random sampling and algorithms like SBC. In this paper, we propose a new BIC based initial training set selection algorithm which can be classified into the latter category. But compared to SBC, it has the following innovations: (1) it uses BIC to judge the status of clusters, (2) it takes different selection strategies for clusters of different status, and the whole selection strategy has taken both representativeness and coverage into consideration.

3. Problem Statements

For a detection system, sometimes events of small probability should be detected. For example, in the audio detection field, for the sake of security, certain audio types belonging to events of small probability are detected to check if there has an incident. Another example is that audiences may be interested in different contents of a film, and so they hope to quickly locate different parts of the film. This requires detecting different audio events in the audio stream. Among these audio events, there inevitably have events of small probability. The labeling workload of small probability events is especially large, because after labeling a very long audio stream, we may have obtained only a small amount of samples of small probability events. So the detection of small probability events demands AL more urgently.

During experiments, we find that when detecting events of small probability, random sampling or SBC would sometimes make AL fail to work or perform poorly since they could not select or select too few samples of small probability events. When taking random sampling to generate the initial training set, due to the small probabilities of small probability events and the random character of random sampling, it is probably that the samples of small probability events would not be selected or only a small amount are selected, and this would cause AL failing to work or performing poorly. The SBC algorithm also has the same problem. Because sometimes the clustering result would not be so desirable, and samples of small probability events would scatter into the clusters that are mainly composed of samples of large probability events, that is to say, there are no clusters that are mainly composed of samples of small probability events. In this case, it is difficult for SBC to select samples of small probability events.

In summary, when detecting events of small probability, random sampling and SBC would make AL fail to work or perform poorly. In order to solve the detection problem of small probability events, we hope that the initial training set would be representative, and in the meantime, would have a good coverage character (the coverage character means that the initial training set should contain enough samples for all classes). Based on the above two criterions, this paper proposes a BIC based initial training set selection algorithm for AL. Experimental results on the "Friends" database and the "daily life" database show that, compared to random sampling and SBC, the proposed BIC based initial training set selection algorithm not only can

efficiently solve the detection problem of small probability events, but also has obvious advantages in detecting events of non-small probability.

4. BIC Based Initial Training Set Selection Algorithm for Active Learning

4.1 Bayesian Information Criterion

Bayesian Information Criterion (BIC) [17] is a model selection criterion. Model selection is to select the model that can best represent a given data set $X = \{x_1, x_2, \ldots, x_N\}$ from some candidate models M_i (the model parameters are θ_i , $I = 1,2, \ldots,m$). The BIC of model M_i is defined as:

$$BIC\left(M_i\right) = \log P\left(x_1, x_2, \cdots, x_N \left| M_i \right.\right) - 1/2\, \lambda \beta_i \log N \quad . (1)$$

$P(x_1, x_2, \ldots, x_N|M_i)$ denotes the maximum likelihood of data set X under model M_i. β_i is the number of independent parameters in parameter set θ_i. Assume that the dimension of the feature vector is d, then β_i is equal to $(d+1/2d(d+1))$. λ is a data-dependent penalty factor (ideally 1.0) to compensate for small sample size cases. The second term $1/2\lambda\beta_i \log N$ is to punish for model complexity. So BIC tends to choose the model that is simple and in the meantime can maximize the value of $BIC(M_i)$. BIC is mostly used in audio segmentation field [18], [19]. In this paper, we expand its application, and use it to judge the status of clusters.

4.2 Judging the Status of Clusters by BIC

Let $X = \{x_i \in R^d, i = 1,2, \ldots, N\}$ denote the sample set of cluster C. Assume that samples of the same class are independent and identically distributed, and can be modeled as one multivariate Gaussian. So if the samples of one cluster belong to the same class, then the cluster can be well modeled by a multivariate Gaussian, that is $X \sim N(\mu,\Sigma)$. In this paper, when performing clustering on the audio database, we find that when the number of clusters is large enough, some clusters are very pure, which means that they are mainly composed of samples belonging to the same class; while the other clusters are mixed clusters, and each mixed cluster is mainly composed of samples coming from two different classes. Such mixed clusters cannot be well modeled by one multivariate Gaussian, but can be better modeled by two Gaussians, one Gaussian for one class.

In each mixed cluster, the two samples that are farthest away from each other can be taken as the two least similar samples within the cluster. These two samples are more likely to come from two different classes. To estimate the status (pure or mixed) of clusters, for each cluster, first, choose the two samples that are farthest away from each other, and then assign the rest samples to these two samples according to nearest-neighbor criterion, thus forming two data sets. If the cluster is a mixed cluster, then these two data sets can be approximately taken as the data sets coming from two different classes. We model each data set by a single Gaussian, then the whole cluster is modeled by two different Gaussians: $N(\mu_1,\Sigma_1)$ and $N(\mu_2,\Sigma_2)$. The judgment of the cluster status can be cast as the following model selection problem:

$$\begin{aligned} M_1 &: X = x_1, x_2, \ldots, x_N \sim N\left(\mu, \Sigma\right) \\ M_2 &: x_1, x_2, \ldots, x_n \sim N\left(\mu_1, \Sigma_1\right); \\ & \quad x_{n+1}, x_{n+2}, \ldots, x_N \sim N\left(\mu_2, \Sigma_2\right). \end{aligned}$$

Model M_1 assumes that all samples within the cluster come from the same class, and can be well modeled by a single Gaussian $N(\mu,\Sigma)$. Model M_2 assumes that n samples (corresponding to one of the two data sets mentioned above) belong to the same class, and can be modeled by a single Gaussian $N(\mu_1,\Sigma_1)$; while the other $(N-n)$ samples (corresponding to the other data set) belong to another class, and can be modeled by another single Gaussian $N(\mu_2,\Sigma_2)$. According to (1), the BIC values of model M_1 and model M_2 can be calculated as follows:

$$\begin{aligned} BIC\left(M_1\right) &= \log P\left(x_1, x_2, \cdots, x_N \left| M_1 \right.\right) - \frac{\lambda}{2}\beta_1 \log N \\ &= \sum_{i=1}^{N} \log P\left(x_i \left| M_1 \right.\right) - \frac{\lambda}{2}\beta_1 \log N \\ &= \sum_{i=1}^{N} \log \frac{1}{\left(2\pi\right)^{d/2} \left| \hat{\Sigma} \right|^{1/2}} \exp\left(-\frac{1}{2}\left(x_i - \hat{\mu} \right)^T \hat{\Sigma}^{-1} \left(x_i - \hat{\mu} \right) \right) - \\ &\quad \frac{\lambda}{2}\beta_1 \log N \\ &= -\frac{d}{2} N \log 2\pi - \frac{N}{2} \log \left| \hat{\Sigma} \right| - \frac{1}{2} \sum_{i=1}^{N} \left(x_i - \hat{\mu} \right)^T \hat{\Sigma}^{-1} \left(x_i - \hat{\mu} \right) - \\ &\quad \frac{\lambda}{2}\beta_1 \log N \\ &= -\frac{d}{2} N \log 2\pi - \frac{N}{2} \log \left| \hat{\Sigma} \right| - \frac{N}{2} - \frac{\lambda}{2}\left(d + \frac{1}{2} d\left(d+1\right) \right) \log N \end{aligned} \quad (2)$$

Similarly, the BIC value of model M_2 is:

$$\begin{aligned} BIC\left(M_2\right) = &-\frac{d}{2} N \log 2\pi - \frac{n}{2} \log \left| \hat{\Sigma}_1 \right| - \\ &\frac{N-n}{2} \log \left| \hat{\Sigma}_2 \right| - \frac{N}{2} - \\ &\lambda\left(d + \frac{1}{2} d\left(d+1\right) \right) \log N \end{aligned} \quad (3)$$

The parameters (μ,Σ), (μ_1,Σ_1), (μ_2,Σ_2) are estimated by corresponding samples, and are denoted as $\left(\hat{\mu},\hat{\Sigma}\right)$, $\left(\hat{\mu}_1,\hat{\Sigma}_1\right)$, $\left(\hat{\mu}_2,\hat{\Sigma}_2\right)$. d is the dimension of the feature vector space. The BIC difference between model M_1 and M_2 can be calculated as a function of the cluster:

$$\begin{aligned}\Delta BIC(C) &= BIC(M_2) - BIC(M_1) \\ &= 1/2\left(N\log|\hat{\Sigma}| - n\log|\hat{\Sigma}_1| - (N-n)\log|\hat{\Sigma}_2|\right) \\ &\quad -1/2\,\lambda\left(d + 1/2\,d(d+1)\right)\log N\end{aligned} \tag{4}$$

According to BIC rule, if $\Delta BIC(C) > 0$, that is, $BIC(M_2) > BIC(M_1)$, which means that modeling cluster C by model M_2 is much better than that by model M_1, then such cluster is more likely to be a mixed cluster. Otherwise, if $\Delta BIC(C) \le 0$, which means that modeling cluster C by model M_1 is much better than that by model M_2, then cluster C is more likely to be a pure cluster.

4.3 Subspace Sample Selection

In order to decrease computational cost, Jiang proposed a subspace sample selection algorithm to reduce the redundancy that exists in samples of the same class [20]. Subspace sample selection selects samples based on the assumption that the sample farther away from the subspace is more difficult to be described by the current subspace. The samples selected by subspace sample selection can generate a subspace which can maximally approach the whole sample space. The procedure of subspace sample selection is:

① Select one sample according to a certain criterion, and put it into the selected sample set.

② Use the selected sample set to generate a subspace.

③ Select the sample farthest away from the subspace, and put it into the selected sample set.

④ Go to ②. The iteration continues until achieving the stopping criterion.

In this paper, we borrow the idea of subspace sample selection, and expand its application. Here we adopt subspace sample selection to select samples in the mixed clusters. Specifically speaking, for a mixed cluster, the selection procedure is:

① Select the sample closest to the cluster centroid as well as the two samples that are farthest away from each other, and put them into the selected sample set.

② Use the selected sample set to generate a subspace.

③ Select the sample that is farthest away from the subspace, and put it into the selected sample set.

④ Go to ②. The iteration continues until a predefined number of samples are selected.

The distance between a sample and the subspace is calculated as follows. Suppose the selected sample set is $X_{sele} = \{x_1, x_2, \ldots, x_s\}$, and the subspace generated by X_{sele} is S_{sele}, then the squared distance between sample x and the subspace S_{sele} is defined as the Minimum Mean-Square Error of using S_{sele} to approach x:

$$\begin{aligned}dist^2(x, S_{sele}) &= \min\left\|x - \sum_{i=1}^{s}\alpha_i x_i\right\|_2^2 \\ &= \min\left\langle\left(x - \sum_{i=1}^{s}\alpha_i x_i\right)\cdot\left(x - \sum_{i=1}^{s}\alpha_i x_i\right)\right\rangle \\ &= \min\left(\langle x\cdot x\rangle - 2\sum_{i=1}^{s}\alpha_i\langle x_i\cdot x\rangle + \sum_{i,j=1}^{s}\alpha_i\alpha_j\langle x_i\cdot x_j\rangle\right)\end{aligned} \tag{5}$$

A real symmetric matrix K can be calculated by the samples in X_{sele}, where

$$K(i,j) = \langle x_i \cdot x_j\rangle\,, \tag{6}$$

Set $$\boldsymbol{K_x} = \left(\langle x_1\cdot x\rangle, \langle x_2\cdot x\rangle, \ldots, \langle x_s\cdot x\rangle\right)^T, \tag{7}$$

$$\boldsymbol{\alpha} = \left(\alpha_1, \alpha_2, \ldots, \alpha_s\right)^T, \tag{8}$$

then formula (5) can be rewritten as :

$$\begin{aligned}dist^2(x, S_{sele}) &= \min\left(\langle x\cdot x\rangle - 2\boldsymbol{K_x}^T\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}\right) \\ &= \min f(\boldsymbol{\alpha})\end{aligned} \tag{9}$$

where $$f(\boldsymbol{\alpha}) = \langle x\cdot x\rangle - 2\boldsymbol{K_x}^T\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}\,. \tag{10}$$

It can be seen that $dist^2(x, S_{sele})$ is the minimum value of function $f(\boldsymbol{\alpha})$. In order to compute the minimum value of $f(\boldsymbol{\alpha})$, set

$$\partial f \big/ \partial\boldsymbol{\alpha} = -2\boldsymbol{K}_x + 2\boldsymbol{K}\boldsymbol{\alpha} = 0\,, \tag{11}$$

then we can obtain:

$$\boldsymbol{\alpha} = \boldsymbol{K}^{-1}\boldsymbol{K}_x\,. \tag{12}$$

Substitute $\boldsymbol{\alpha} = K^{-1}K_x$ for $\boldsymbol{\alpha}$ in (9), and then we can get:

$$dist^2(x, S_{sele}) = \langle x\cdot x\rangle - \boldsymbol{K_x}^T\boldsymbol{K}^{-1}\boldsymbol{K_x}\,. \tag{13}$$

4.4 Specific Strategies of the BIC Based Initial Training Set Selection Algorithm

The specific strategies of the BIC based initial training set selection algorithm are:

① Perform clustering on the whole unlabeled training set, and obtain clusters C_i, $i = 1,2, \ldots, N$.

Here, the k-means clustering algorithm which is adopted by SBC is used.

② For each cluster C_i, select samples according to its status as follows.

1) $X_{sele} = X_{sele} \cup centroid(C_i)$, if $|C_i| < Th$

$centroid(\cdot)$ denotes the sample closest to the cluster centroid. $|\cdot|$ denotes the number of samples contained in the cluster. For a cluster, if its size is less than the threshold

Th, then take it as a small cluster, otherwise a large cluster. When a cluster is small, the calculated BIC value would not be accurate enough, and then the ΔBIC value cannot well describe its status. So for a small cluster, we do not calculate its ΔBIC, but just select the sample closest to the cluster centroid as the representative sample.

2) $X_{sele} = X_{sele} \cup centroid(C_i)$, if $|C_i| \geq Th$ and $\Delta BIC(C_i) \leq 0$

For a large cluster, calculate its ΔBIC value according to Section 4.2. If $\Delta BIC \leq 0$, it illustrates that this cluster is more likely to be a pure cluster. For a pure cluster, we mainly consider the representativeness. So the sample closest to the cluster centroid is selected.

3) $X_{sele} = X_{sele} \cup sub(C_i)$, if $|C_i| \geq Th$ and $\Delta BIC(C_i) > 0$

$sub(\cdot)$ denotes the samples selected by subspace sample selection which is described in Section 4.3. For a large cluster, if $\Delta BIC > 0$, it illustrates that this cluster is more likely to be a mixed cluster. For a mixed cluster, only the cluster centroid is not enough. The coverage character should be given more consideration, otherwise, samples of certain classes could not be selected at all, especially for those classes of small probability events. So just as that described in Section 4.3, for the mixed cluster, besides the sample closest to the cluster centroid and the two samples that are farthest away from each other, another several samples are selected by subspace sample selection. The selected number is proportional to the size of the cluster.

During clustering experiments, we find that sometimes the samples of small probability events would scatter into the clusters that are mainly composed of samples of non-small probability events, thus forming some mixed clusters. In this case, the centroid of the mixed cluster is generally a sample of non-small probability events. So only selecting cluster centroids is not enough, for the samples of small probability events would not be selected. When selecting samples in such a mixed cluster, if the current selected sample set doesn't contain samples of small probability events, then samples of small probability events would be far away from the subspace generated by the current selected sample set. Since subspace sample selection selects the sample that is farthest away from the subspace in each iteration, it can well select samples of small probability events. Also since the cluster centroid is first selected to augment the selected sample set, and the cluster centroid is less likely to be a class boundary sample, then the class boundary samples would be far away from the subspace generated by the selected sample set, which means that the subspace sample selection algorithm can select class boundary samples. This is very beneficial to model training, especially for those discriminant models. The class boundary samples can provide the model a higher initial performance, and a good initial performance can offer AL a good starting point.

It can be seen that our BIC based initial training set selection algorithm has taken both representativeness and coverage into consideration. The framework of the BIC based initial training set selection algorithm is shown in Fig. 1.

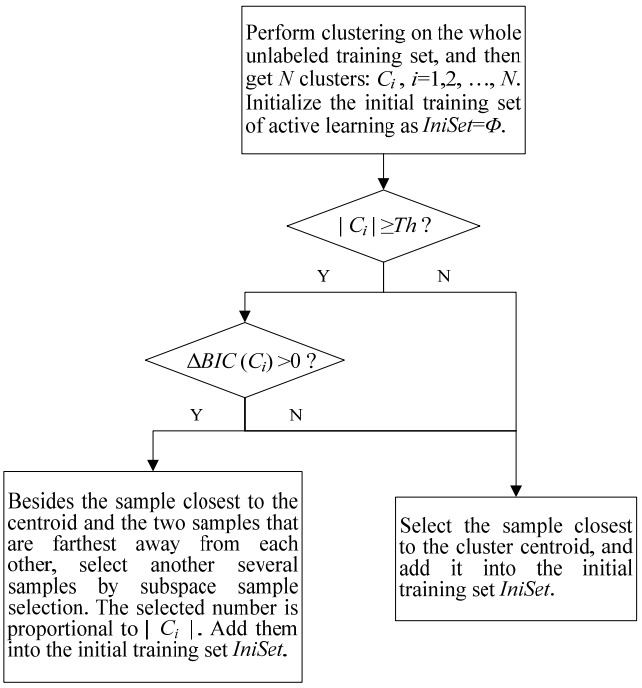


Fig.1. The framework of the BIC based initial training set selection algorithm.

5. Experimental Results and Analysis

5.1 Experiments on Toy Data Set

To visually verify the feasibility of judging the status of clusters by BIC, we first do experiments on a toy data set in two-dimensional space. There are two classes in the toy data set, and the two classes obey the normal distribution of $N(\mu_1,\Sigma)$ and $N(\mu_2,\Sigma)$ respectively, where $\mu_1=[2,2]^T$, $\mu_2=[4,4]^T$ and $\Sigma=[1,1]^T$. Each class contains 200 samples. Set the number of clusters to 3, and take k-means to cluster all the samples in the toy data set. The clustering result is shown in Fig.2. Symbols "o" and "*" are used to distinguish the two classes. The three clusters C_1, C_2 and C_3 are marked black, blue and red respectively.

From Fig. 2 we can see that C_1 and C_2 are two much pure clusters, while C_3 is obviously a mixed cluster. Tab. 1. shows the ΔBIC values of the three clusters with $\lambda = 6$. It can be seen that the ΔBIC values of C_1 and C_2 are less than zero, while the ΔBIC value of the mixed cluster C_3 is larger than zero, so it is feasible to judge the status of clusters by BIC.

Clusters	C_1 (black)	C_2 (blue)	C_3 (red)
ΔBIC	-7.6816	-6.9676	19.866

Tab. 1. The ΔBIC values of the three clusters.

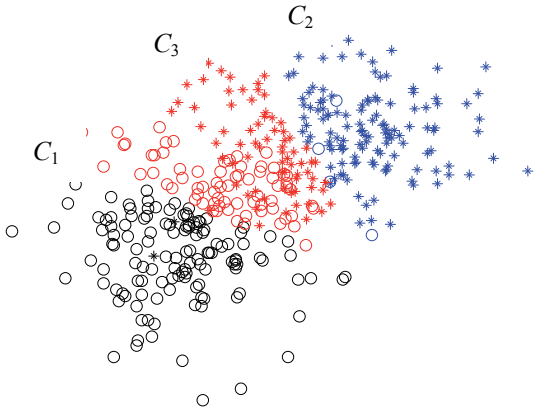


Fig. 2. The clustering result of the toy data set. The resulting three clusters C_1, C_2 and C_3 are marked black, blue and red respectively.

5.2 Experiments on Real Data Set

(1) Experimental Setting

In this section, experiments are done on two real data sets to verify the effectiveness of the proposed BIC based initial training set selection algorithm. The first dataset is constructed by 10 episodes of melodrama "Friends". The extracted audio tracks are totally about 3.68 h in length, and the contained audio events are usually distinct enough to be perceived. 6 major semantic classes that occur in every of the 10 episodes have been labeled, including speech, laugh, music, silence, applause and door-close. For segments in which several audio events are mixed, we only label the dominant one. The severely mixed segments for which it is difficult to recognize the exact semantic, and the segments with the semantic classes that only occasionally occur in one or two of the 10 episodes are all labeled as others. The second data set which we call "daily life" data set is constructed by a 2.75-hour-long audio document recorded by a mobile phone. This audio document has recorded the sounds of a student's daily life. The audio segments are labeled as one of the 6 semantic classes that describe the audio scenes: classroom, music, lab, playground, station and water-room. The segments labeled as classroom mainly contain speaking of a teacher and several students. Segments labeled as music contain musical sound of different styles. Segments labeled as lab mainly contain sounds of typing on the keyboard, clicking of the mouse and footsteps. Segments labeled as playground mainly contain sounds of the basketball hitting the ground and the shouting of students. Segments labeled as station mainly contain sounds of the bus driving, braking and launching. Segments labeled as water-room mainly contain the sound of water flowing.

The audio documents are in mono channel format, and are down-sampled to 16 kHZ. The frame length\ shift is 30\10 ms. For each frame, a set of features are extracted, including short-time energy, zero crossing rate, 8 dimensional Mel Frequency Cepstral Coefficients, sub-band spectral flux, brightness and bandwidth. Then the audio documents are redivided into adjacent clips of one second long with 50% overlap. The means and standard deviations of the above features are computed over all frames contained in one clip. For each clip, the following long time features: high zero crossing rate, low energy rate and spectrum flux are also extracted, thus forming a 43 dimensional feature vector for each clip. The clips are taken as the basic detecting units. For the "Friends" data set, 7 episodes are randomly chosen to construct the training set, and the remaining 3 the test set. The training set is about 2.55 h in length, of which speech, laugh, music, silence, applause and door-close occupy 53.74%, 21.57%, 14.73%, 1.6%, 0.4% and 1.18%, respectively. The test set is about 1.13 h in length, of which the above 6 semantic classes occupy 41.48%, 18.29%, 8.52%, 2.76%, 0.74% and 1.11%, respectively. Whether in the training set or in the test set, the proportions of silence, applause and door-close are all lower than 3%, so these three audio events are obviously the events of small probability. For the "daily life" data set, 70% of the total samples are randomly selected to construct the training set, and the remainder the test set. The training set is about 1.91 h in length, of which classroom, music, lab, playground, station and water-room occupy 24%, 18.18%, 30.55%, 18.18%, 5.82% and 3.27%, respectively. The test set is about 0.84 h in length, of which the above 6 semantic classes occupy 24.79%, 18.18%, 29.75%, 18.18%, 6.61% and 2.48%, respectively. The water-room class accounts for a small proportion in the audio document. It can be taken as an event of small probability.

The size of the initial training set can be determined according to the acceptability of the labeling workload. In this paper, we set it to be 3% of the size of the whole training set. The rest 97% serves as the unlabeled training set. Since the initial training samples are few, SVM [21] is adopted as the classifier, as it is one of the most competitive classifier for small samples problem. A more theoretical consideration is given in [22]. For SVM, the most popular AL algorithm is the one proposed by Tong&Koller [3], denoted as SVM_{AL}. SVM_{AL} selects the sample closest to the hyperplane in each iteration for manual labeling. Readers can refer to [3] for its detailed principles. The kernel function used is the radial basis function, and the two SVM parameters, C and γ, are selected based on 5-fold cross-validation. Since SVM is a binary classifier, the one-vs-all binary classification technology is adopted to detect a certain audio event in the continuous audio stream. So the samples of the audio event that we want to detect should be relabeled as positive, and all the other samples negative. The detecting procedure is actually a binary classification procedure which is to label the clips of the audio stream as being the audio event of people's interest or not.

To verify the effectiveness of the proposed BIC based initial training set selection algorithm, here we do experiments to compare it with random sampling and SBC [13]. To be fair, the size of the initial training set should be iden-

tical for the three algorithms. For random sampling, it is easy to control the selected number; while for SBC, since it selects only one sample in each cluster, the selected number would be less than that of the BIC based algorithm under the same clustering result. Thus, for SBC, we take the following two strategies to ensure that its selected number is identical with that of the BIC based algorithm: 1) In order to compare SBC with the BIC based algorithm under the same clustering result, for each small cluster (the cluster whose size is less than the threshold *Th*), select the sample closest to the cluster centroid; while for each large cluster, select p-nearest neighbors of the centroid. p is proportional to the size of the cluster. Denote this strategy as SBC-A. 2) Enlarge the number of clusters to the predefined size of the initial training set, just as the literature [13] did, and then re-cluster. For each cluster, select the sample closest to the cluster centroid. Denote this strategy as SBC-B.

All experiments are run 10 times, and the average is taken as the final result. To comprehensively evaluate the detecting precision rate and recall rate, take F1 measure as the evaluation criterion:

$$F1 = \frac{2 \times recall \times precision}{\left(recall + precision\right)} \times 100\% \tag{14}$$

where the detecting precision rate is defined as:

$$precision = \frac{\begin{array}{c}the\ number\ of\\ correctly\ detected\ samples\ of\ a\ certain\ event\end{array}}{\begin{array}{c}the\ number\ of\\ all\ samples\ that\ are\ recognized\ as\ a\ certain\ event\end{array}} \times 100\% \tag{15}$$

and the detecting recall rate is defined as:

$$recall = \frac{\begin{array}{c}the\ number\ of\\ correctly\ detected\ samples\ of\ a\ certain\ event\end{array}}{\begin{array}{c}the\ total\ number\ of\\ samples\ of\ a\ certain\ event\ in\ the\ database\end{array}} \times 100\% \tag{16}$$

(2) Experimental Results on Detecting Events of Small Probability

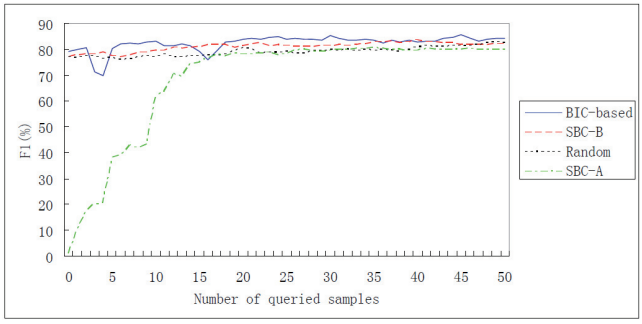
To verify the effectiveness of the proposed BIC based initial training set selection algorithm in detecting events of small probability, here we adopt SVM_{AL} to detect silence, applause and door-close in the "Friends" data set respectively, and to detect water-room in the "daily life" data set. Compare the performances of SVM_{AL} under the four different initial training sets obtained by random sampling, SBC-A, SBC-B and the proposed BIC based algorithm.

Fig. 3, 4, 5 show under the four different initial training sets, the F1-scores over the number of queries when detecting silence, applause and door-close in the "Friends" data set respectively, and Fig. 6 shows the result of detecting water-room in the "daily life" data set. In order to show the result clearly and comprehensively, in each figure, we provide two subfigures. The subfigure (a) shows the average of the 10 independent experiments. Besides the average, the subfigure (b) also shows the standard error indicated by the error bar. It should be noticed that in the subfigure (b) of Fig. 3 and Fig. 4, including the result of SBC-A would cause the figure to be unclear and difficult to read, so we have removed it.

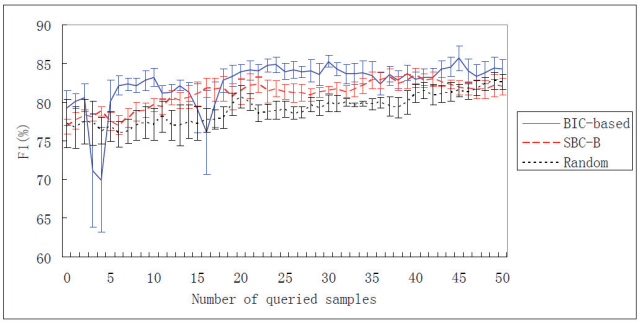
The number of clusters is set to 30 for SBC-A and the BIC based algorithm since after clustering the training samples into 30 clusters, the pure clusters are pure enough, and the mixed clusters are mainly composed of samples coming from two different classes. In practice, since the class label of each sample is unknown in advance, in order to determine the number of clusters, we can repeatedly set the cluster number to a much larger value, and then after clustering, randomly select a small number of samples in each cluster to see whether the clustering result is satisfying. As the literature [13] did, the number of clusters is set to the predefined size of the initial training set (3% of the size of the whole training set in this paper) for SBC-B. Set $\lambda = 0.14$ for the "Friends" data set, and $\lambda = 0.2$ for the "daily life" data set (the setting of λ is discussed in subsequent (4)). Set $Th = 60$. The Th value can be determined as follows:

As described in Section 4.2, to estimate the status of clusters, the samples in the cluster would be grouped into two data sets. If the sizes of the data sets are too small, the BIC is not applicable, because the BIC value is calculated based on the assumption that the data set can be modeled by a Gaussian, while when the data set is too small, it cannot be modeled by a Gaussian. So after grouping the samples in the cluster into two data sets, collect all the clusters in which the size of the smaller data set is less than the threshold *th* (*th* is set according to the empirical value of BIC), and then the smallest size of these clusters can be taken as a reference of setting *Th*.

For both the "Friends" data set (Fig. 3, 4, 5) and the "daily life" data set (Fig. 6), when selecting the initial training samples, the proposed BIC based algorithm has selected all the semantic classes in each running. Random sampling, SBC-A and SBC-B have selected all the semantic classes in each running in the "daily life" data set, but sometimes fail to select samples of small probability events in the "Friends" data set. When the samples of a certain event are failed to be selected, then SVM_{AL} cannot be carried out in detecting this event. In this case, record the F1 values as zeros. So compared to random sampling, SBC-A and SBC-B, the BIC based algorithm can solve the detection problem of small probability events, for it can effectively select the samples of small probability events. For random sampling, because of its randomness character, sometimes it would fail to select the samples of small probability events. For SBC-B, if the clustering result is not optimal, that is, if there are no clusters of small probability events, then the samples of small probability events are

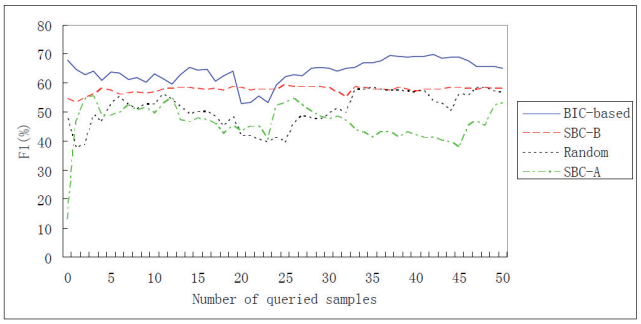


(a)

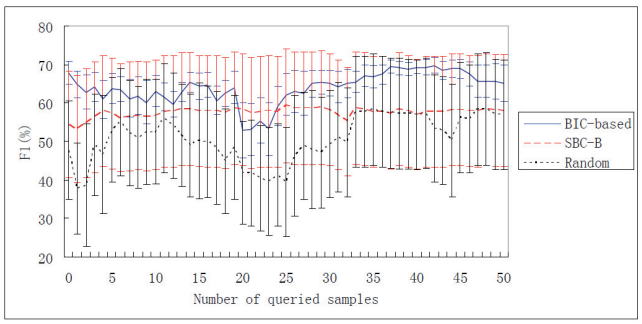


(b)

Fig. 3. Under the four different initial training sets, the F1-scores over the number of queries when detecting silence in the "Friends" data set. Subfigure (a) shows the average of the 10 independent experiments. Subfigure 2(b) adopts error bar to show the standard error.

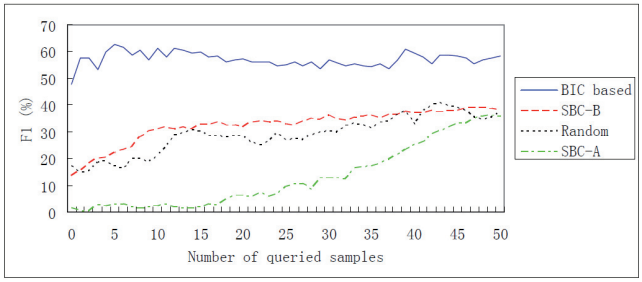


(a)

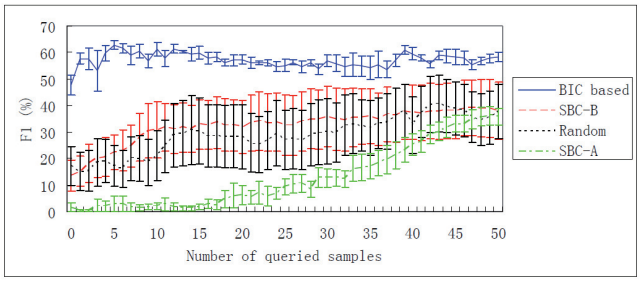


(b)

Fig. 4. Under the four different initial training sets, the F1-scores over the number of queries when detecting applause in the "Friends" data set. Subfigure (a) shows the average of the 10 independent experiments. Subfigure (b) adopts error bar to show the standard error.

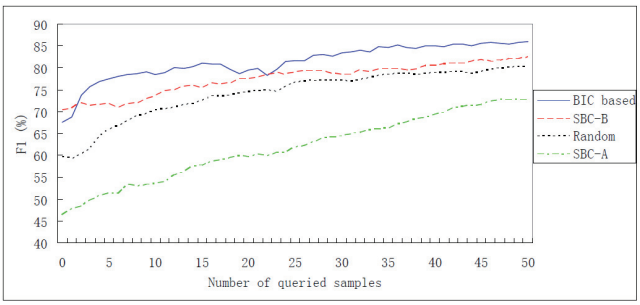


(a)

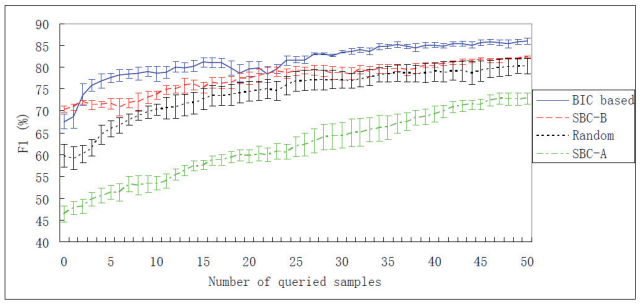


(b)

Fig. 5. Under the four different initial training sets, the F1-scores over the number of queries when detecting door-close in the "Friends" data set. Subfigure (a) shows the average of the 10 independent experiments. Subfigure (b) adopts error bar to show the standard error.



(a)



(b)

Fig. 6. Under the four different initial training sets, the F1-scores over the number of queries when detecting water-room in the "daily life" data set. Subfigure (a) shows the average of the 10 independent experiments. Subfigure (b) adopts error bar to show the standard error.

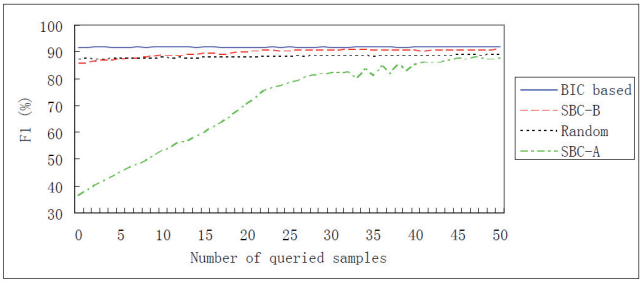
probably not to be selected, because the sample closest to the cluster centroid is less likely to be a sample of small probability events. For SBC-A, the reason is much the same as that of SBC-B.

It can be seen from Fig. 3 to Fig. 6 that SVM_{AL} performs best under the initial training set obtained by the BIC based algorithm, while performs worst under the initial training set obtained by SBC-A. Just as discussed in Section 4.4, the BIC based algorithm can select certain class boundary samples since it adopts subspace sample selection to select samples in the mixed cluster. Such boundary samples are very informative to the discriminant model SVM. Thus the BIC based algorithm can offer the active learning algorithm SVM_{AL} a good starting point. Also when selecting the initial training samples, the BIC based algorithm has taken both representativeness and coverage into consideration. Due to the above reasons, the BIC based algorithm can perform better than the other two algorithms, random sampling and SBC. For random sampling, due to its randomness character and the small size of the initial training set, its selected sample set cannot well describe the whole training set, so it performs much worse. For SBC-A, since it selects several samples close to the cluster centroid in the large clusters, its selected sample set would be redundant and less representative. Also its selected samples are less likely to be on the class boundary. Maybe because of the above reasons, SBC-A performs worst. For SBC-B, its selected sample set is representative, so it performs better than SBC-A. But it has not considered the coverage character, so sometimes it would fail to select samples of small probability events. In this case, when detecting such a small probability event, its F1-measure is set to zero. Also since it selects the sample closest to the cluster centroid in each cluster, its selected samples are less likely to be on the class boundary. Due to the above reasons, SBC-B performs worse than the BIC based algorithm.

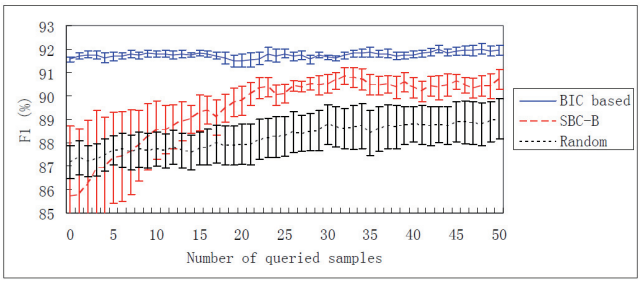
(3) Experimental Results on Detecting Events of Non-small Probability

Besides verifying the effectiveness of the BIC based algorithm in detecting events of small probability, we also do experiments to investigate its superiority in detecting events of non-small probability. Here we use SVM_{AL} to detect speech in the "Friends" data set, and to detect lab in the "daily life" data set. These two audio classes appear most frequently in the two data sets respectively. Fig. 7 and Fig. 8 show under the four different initial training sets, the F1-scores over the number of queries when detecting speech in the "Friends" data set and detecting lab in the "daily life" data set respectively.

It is easy for the BIC based algorithm, random sampling, SBC-A and SBC-B to select samples of non-small probability events, so in both data sets, they have all selected the classes of non-small probability events in each running. From Fig. 7 and Fig. 8 it can be seen that when detecting events of non-small probability, it is still under the initial training set obtained by the BIC based algorithm that SVM_{AL} performs best. It is because that the BIC based algorithm has taken both representativeness and coverage into consideration, then its selected samples of non-small probability events are representative, and its selected samples can better describe the whole training set compared to

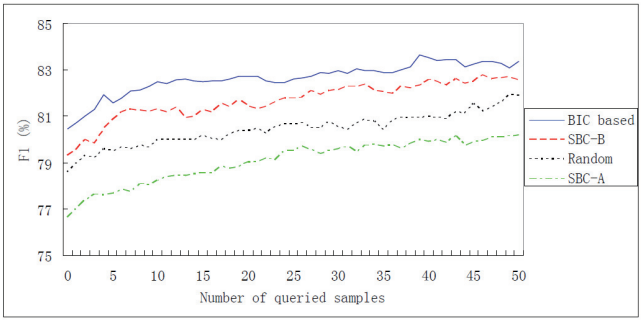


(a)

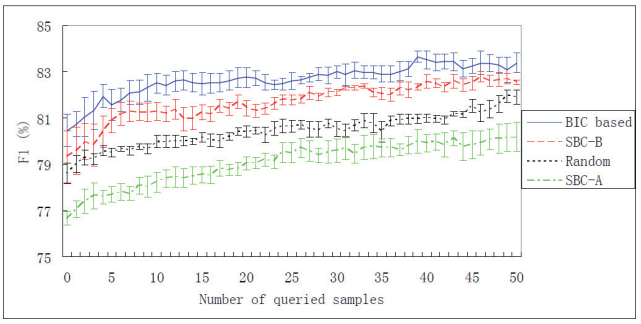


(b)

Fig. 7. Under the four different initial training sets, the F1-scores over the number of queries when detecting speech in the "Friends" data set. Subfigure (a) shows the average of the 10 independent experiments. Subfigure (b) adopts error bar to show the standard error.



(a)



(b)

Fig. 8. Under the four different initial training sets, the F1-scores over the number of queries when detecting lab in the "daily life" data set. Subfigure (a) shows the average of the 10 independent experiments. Subfigure (b) adopts error bar to show the standard error.

random sampling, SBC-A and SBC-B. SBC-A and SBC-B have considered the representativeness but not the coverage character, and then their selected sample sets describe the whole training set worse than that of the BIC based algorithm, so they perform worse than the BIC based algorithm. Compared to SBC-A, SBC-B would have more advantages. Because SBC-B enlarges the number of clusters, then the clusters would be much pure, and the sample closest to the cluster centroid would be more representative.

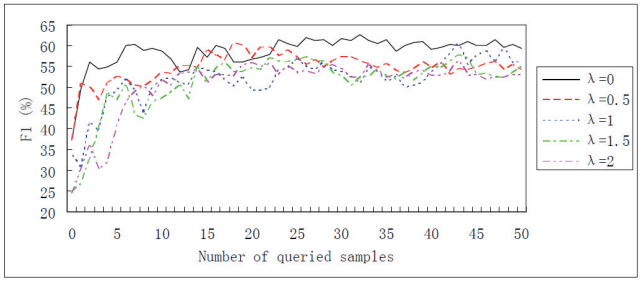
In summary, the proposed BIC based initial training set selection algorithm can effectively select samples of small probability events, thus can solve the detection problem of small probability events. It shows obvious advantages both in detecting events of small probability and in detecting events of non-small probability. When the number of clusters is equal to the size of the initial training set, SBC is much effective in detecting events of non-small probability. But sometimes it would fail to select samples of small probability events, and then performs much worse in detecting events of small probability. Whether in detecting events of small probability or in detecting events of non-small probability, random sampling performs worse than the BIC based algorithm and SBC (when the number of clusters is equal to the size of the initial training set).

(4) The Penalty Factor λ

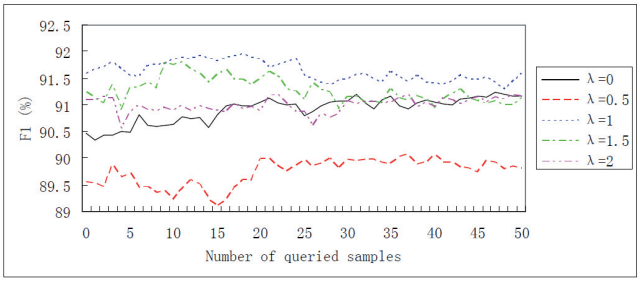
It can be seen from (4) that when using the BIC based algorithm to construct the initial training set, the penalty factor λ should be determined in advance in order to calculate ΔBIC. Currently, the determination of λ is still a difficulty. Some researchers propose that its theoretical value is 1, but $\lambda = 1$ cannot always give satisfactory results in practical application. Since the theoretical value of λ is 1, in this paper we take different values around 1 for λ, and determine its final value through repeated experiments.

In this section, we do experiments on the "Friends" data set to show the relationship between the detection performance and the parameter λ. Here we take the following two experimental schemes. 1) When λ takes different values, for the mixed clusters, keep the sampling rate unchanged. Thus under different values of λ, the sizes of the initial training sets are different. 2) When λ takes different values, keep the size of the initial training set unchanged. Thus under different values of λ, the sampling rates of the mixed clusters are different.

If λ is too large, from (4) we can see that most clusters' ΔBIC value would be less than zero, which means that certain mixed clusters would be wrongly recognized as pure clusters. Since for pure clusters, the sample closest to the cluster centroid is selected, the larger the λ is, the more likely that the BIC based algorithm would degenerate into SBC. So λ should not be too large. In this paper, we set λ to be equal to 0, 0.5, 1, 1.5 and 2 respectively, and show the relationship between the detection performance and the parameter λ in Fig. 9 and Fig. 10.

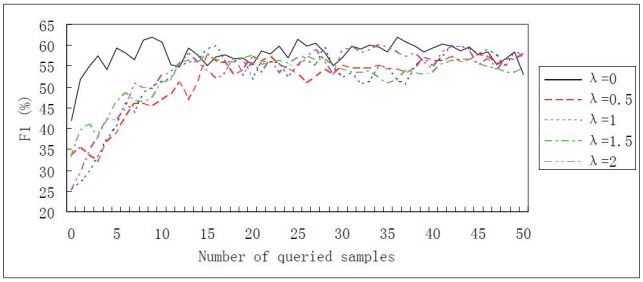


(a)

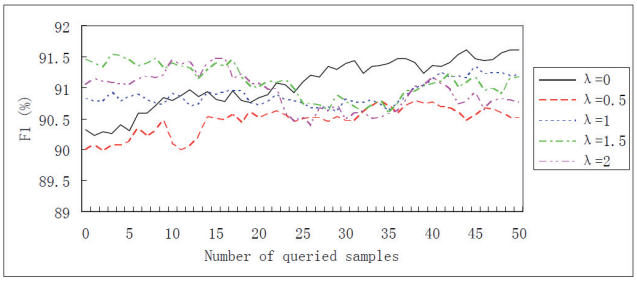


(b)

Fig. 9. The performance comparison of SVM_{AL} under different values of λ when keeping the sampling rate unchanged. Subfigure (a) is the result of detecting door-close and subfigure (b) is the result of detecting speech.



(a)



(b)

Fig. 10. The performance comparison of SVM_{AL} under different values of λ when keeping the size of the initial training set unchanged. Subfigure (a) is the result of detecting door-close and subfigure (b) is the result of detecting speech.

1) When λ takes different values, keep the sampling rate unchanged

Fig. 9 shows when keeping the sampling rate unchanged, the performance comparison of SVM_{AL} under different values of λ. Subfigure (a) is the result of detecting door-close and subfigure (b) is the result of detecting

speech. It can be seen that when detecting the event of small probability – door-close, smaller λ can get a better result. While when detecting the event of non-small probability – speech, $\lambda = 1$ can get the best result. The reasons are discussed as follows.

According to (4), the smaller the λ is, the larger the ΔBIC value would be. So smaller λ would cause more clusters possessing a ΔBIC value larger than zero. This is very beneficial for selecting samples of small probability events. Because more clusters possessing a ΔBIC value larger than zero means that there are more mixed clusters, and certain pure clusters would be wrongly recognized as mixed clusters. For the mixed cluster, the subspace sample selection is adopted to select samples, and the subspace sample selection algorithm can well select samples of small probability events as that discussed in Section 4.4. Therefore, smaller λ is helpful in detecting events of small probability.

However, the value of λ is not the smaller the better. Because small λ would cause certain pure clusters to be wrongly recognized as mixed clusters. Taking the subspace sample selection to select samples in such wrongly recognized pure clusters would introduce redundancy and then reduce the representativeness of the selected samples. Large λ would cause certain mixed clusters to be wrongly recognized as pure clusters. For such wrongly recognized mixed clusters, only the sample closest to the cluster centroid is selected, and this would cause the initial training set having a worse coverage character. Worse coverage means worse description of the whole sample space. That is why when detecting speech, neither larger λ (> 1) nor smaller λ (< 1) could get a better result than that with $\lambda = 1$.

On the whole, smaller λ would produce better performance in detecting events of small probability. However, when detecting events of non-small probability, the performance difference between using larger λ and using smaller λ is not so large. So taking the detection of small probability events into consideration, the coverage character should be given more considerations. That is to say, λ should take a smaller value. In this paper, repeated experiments are done to further adjust the parameter λ in the range of 0~0.5, and finally $\lambda = 0.14$ and $\lambda = 0.2$ are set for the "Friends" data set and for the "daily life" data set respectively.

2) When λ takes different values, keep the size of the initial training set unchanged

In 1), when λ takes different values, keeping the sampling rate unchanged could cause the initial training sets having different sizes. To be fair, we hope to investigate the relationship between the detection performance and the parameter λ under the same initial labeling workload, and then we set the size of the initial training set to be 3% of the size of the whole training set. Fig. 10 shows the performance comparison of SVM_{AL} under different values of λ when keeping the size of the initial training set unchanged. Subfigure (a) is the result of detecting door-close and subfigure (b) is the result of detecting speech.

It can be seen from Fig. 10 that when detecting the event of small probability – door-close, due to the reasons discussed in 1), SVM_{AL} performs best with $\lambda = 0$, and performs much worse in the first few iterations when λ takes the other four values. When detecting the event of non-small probability – speech, larger λ ($\lambda = 1$, 1.5, 2) produces a better performance in the first few iterations, but the whole performance is very unstable. Just as discussed in 1), the reason may be that larger λ has caused certain mixed clusters to be wrongly recognized as pure ones, and this would cause a poor coverage of the whole sample space. With a smaller λ ($\lambda = 0$), SVM_{AL} performs slightly worse in the first few iterations, but its performance increases rapidly with the increasing of iterations. Moreover, the whole performance has small fluctuations. So under the same initial labeling workload, still small λ would be suitable for constructing a good initial training set.

In summary, the relationship between the detection performance and the parameter λ tells us that when there exist events of small probability, small λ (usually smaller than 1) would be proper. This conclusion would help to narrow the search range of the optimal λ.

6. Conclusions

In this paper, we propose a BIC based initial training set selection algorithm to solve the detection problem of small probability events. The innovations of our BIC based algorithm are: (1) it uses BIC to judge the status of clusters; (2) it takes different selection strategies for clusters of different status, and the whole selection strategy has taken both representativeness and coverage into consideration. The experimental results demonstrate: (1) it is feasible to judge the cluster status by BIC; (2) the proposed BIC based initial training set selection algorithm can effectively solve the detection problem of small probability events; (3) the BIC based algorithm also shows obvious advantages in detecting events of non-small probability. When calculating ΔBIC, the determination of λ is still a difficulty. In this paper, we simply determine it through repeated experiments. In future work, a more effective determination method should be studied.

Acknowledgements

This work is partially supported by the Project of Shandong Province Higher Educational Science and Technology Program (No. J12LN23), Research Fund for Excellent Young and Middle-aged Scientists of Shandong Province (No. BS2012DX038), China Postdoctoral Science Foundation (No. 2012M511538), Post-doctoral Innovation Fund of Shandong Province(201202032), National Natural

Science Foundation of China (No. 61201441), and Ji'nan City University Independent Innovation Program (No. 201202018).

References

[1] COHN, D. A., GHAHRAMANI, Z. B., JORDAN, M. I. Active learning with statistical models. In *Proceedings of Neural Information Processing Systems,* 1994, p. 705 - 712.

[2] REZAEE, P., TAYARANI, M., KNOECHEL, R. Miniaturized microstrip filter design using active learning method. *Radioengineering*, 2011, vol. 20, no. 4, p. 857 – 865.

[3] TONG, S., KOLLER, D. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2001, vol. 2, p. 45 – 66.

[4] DEMIR, B., PERSELLO, C., BRUZZONE, L. Batch-mode active-learning methods for the interactive classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing,* 2011, vol. 49, no. 3, p. 1014 – 1031.

[5] NGUYEN, H. T., SMEULDERS, A. W. M. Active learning using pre-clustering. In *Proceedings of International Conference on Machine Learning*. 2004, p. 623 – 630.

[6] BEYGELZIMER, A., DASGUPTA, S., LANGFORD, J. Importance weighted active learning. In *Proceedings of International Conference on Machine Learning*. 2009, p. 7 – 56.

[7] HE, X. F. Laplacian regularized D-optimal design for active learning and its application to image retrieval. *IEEE Transactions on Image Processing*, 2010, vol. 19, no. 1, p. 254 – 263.

[8] SETTLES, B. Active learning literature survey. *Computer Sciences Technical Report 1648*. University of Wisconsin–Madison, 2010.

[9] ZHU, J. B., WANG, H. Z., TSOU, B. K., et al. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech & Language Processing*, 2010, vol. 18, no. 6, p. 1323 – 1331.

[10] SEUNG, H. S., OPPER, M., SOMPOLINSKY, H. Query by committee. In *Proceedings of 5th Annual Workshop on Computational Learning Theory*, 1992, p. 287 - 294.

[11] YE Z. X., BERGER, T. *Information Measures for Discrete Random Fields*. Beijing and New York: Science Press, 1998.

[12] HU, R., NAMEE, B. M., DELANY, S. J. Off to a good start: using clustering to select the initial training set in active learning. In *Proceedings of the Florida AI Research Society Conference*. 2010, p. 26 - 31.

[13] ZHU, J. B., WANG, H. Z., YAO, T. S., 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*. 2008, p. 1137 - 1144.

[14] KANG, J., RYU, K. R., KWON, H. C. Using cluster-based sampling to select initial training set for active learning in text classification. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2004, p. 384 - 388.

[15] YUAN W. W., HAN, Y. K., GUAN, D. H., et al. Initial training data selection for active learning. In *Proceedings of International Conference on Ubiquitous Information Management and Communication*. 2011, p. 1 - 7.

[16] CEBRON, N., BERTHOLD, M. R. Adaptive active classification of cell assay images. In *Proceedings of Principles of Data Mining and Knowledge Discovery*. 2006, p. 79 - 90.

[17] ŽĎÁNSKÝ, J. Detection of acoustic change-points in audio streams and signal segmentation. *Radioengineering*, 2005, vol. 14, no. 1, p. 37 – 40.

[18] XUE, H., LI, H. F., GAO, C., et al. Computationally efficient audio segmentation through a multi-stage BIC approach. In *Proceedings of 3rd International Congress on Image and Signal Processing*. 2010, p. 3774 – 3777.

[19] CHENG, S. S., WANG, H. M., FU, H. C. BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. *IEEE Transactions on Audio, Speech & Language Processing*, 2010, vol. 18, no. 1, p. 141 – 157.

[20] JIANG, W. H., ZHOU, X. F., YANG, J. Y. Subspace sample selection for SVM on face recognition (in Chinese). *Computer Engineering and Application*, 2007, vol. 43, no. 20, p. 14 – 17.

[21] WANG, L. *Support Vector Machines: Theory and Applications*. Springer, 2005.

[22] ZHANG, T., OLES, F. J. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of International Conference on Machine Learning*. 2000, p. 1191 to 1198.

About Authors ...

Yan LENG was born in Yantai, China, in 1981. She received both the B.E. degree and the M.E. degree from Shandong University (SDU), Ji'nan, China in 2003 and 2006 respectively and received the D.E. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China in 2012. Her research interests include audio classification, audio detection, audio retrieval and medical image processing.

Guang-hui QI was born in Heze, China, in 1976. He received the B.E. degree from Shandong University of Technology (SDUT), Ji'nan, China in 1999, and received the M.E. degree from Shandong University (SDU), Ji'nan, China in 2001. His research interests include audio feature extraction and audio classification.

Xin-yan XU was born in Ji'nan, China, in 1962. She received the M.E. degree from Shandong University (SDU), Ji'nan, China in 2006. Her research interests include audio classification and medical image processing.