A Multicriterion Query-Based Batch Mode Active Learning Technique

Yang Jiao, Pengpeng Zhao, Jian Wu, Yujie Shi and Zhiming Cui

Abstract Active learning is well-motivated in many modern machine learning problems, where unlabeled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain. The selection and use of sampling strategy is core of active learning. Most active learning methods select those uncertain or representative unlabeled samples to query their labels. And some of the active learning methods consider both in the query selection. However, the uncertainty sampling methods rely on the relative correctness or confidence of the current model and suffer from a lack of the feature space. The representative sampling methods avoid the drawbacks associated with uncertainty sampling, but tend not to improve the learning model very efficiently. The combining methods are lack of the consideration of sample redundancy. This paper proposes a multicriterion active learning technique for solving multiclass problems. First, use the Best-versus-Second-Best (BvSB) method to calculate the sample's uncertainty and then select the most valuable component to constitute the uncertain set; further, use the kernel *k*-means clustering algorithm and the resulting sample set is divided into *h* different clusters; finally, use Gaussian process to select the most informative sample in each cluster and submit to human experts for annotation. The results show that the labeling cost can be reduced without degrading the performance.

Keywords Active learning · Sampling strategy · Uncertainty · Representative · Diversity

Y. Jiao · P. Zhao (✉) · J. Wu · Y. Shi · Z. Cui
Department of Computer Science and Technology, Soochow University, Suzhou 215006, China
e-mail: ppzhao@vip.sina.com

Z. Wen and T. Li (eds.), *Foundations of Intelligent Systems*,
Advances in Intelligent Systems and Computing 277,
DOI: 10.1007/978-3-642-54924-3_63,

1 Introduction

Many supervised learning algorithms of the machine learning area have been widely used in pattern recognition tasks. In all of these methods, the accuracy of classifier depends heavily on the labeled sample set. However, in reality, to obtain the training samples is very easy while the labeled samples are scarce, so to get labeled samples requires a high price. Moreover, redundant-labeled samples may slowdown the training speed, while they are not helpful to the classifier. In order to reduce the time and cost of labeling [1], which requires that the sample selected during training, not only has the information content but be diversed from each other. Active learning [2] is an effective method to solve these problems. Active learning algorithms select high-information content unlabeled examples to be labeled by experts [3, 4], several loops make the correct classification accuracy gradually increased, and thus the classifier obtains the strong generalization ability in the case of minimum labeling cost. Compared with the traditional supervised learning methods, active learning can significantly reduce time and cost of labeling. How to choose samples, as few as possible to get the higher classification accuracy, are the core issue of active learning algorithm. Therefore, the sampling strategy [5] naturally becomes a concern of active learning algorithms.

The traditional sampling methods are generally divided into two categories: one is based on uncertainty [6], merely use the uncertainty to measure sample information content, select the most uncertain samples of classification for labeling; although this method has a wide range of applications and achieved good results in practical tasks, but it only considers the relationship between current sample and the labeled samples, ignoring the distribution information of unlabeled sample set. Thus the important issue is that the unavoidable choice of outliers in the training process may reduce the classification accuracy. Another method surmounts the difficulty of uncertainty sampling, considering the sample of uncertainty and representativeness [7]. But for different data, it is difficult to measure the importance level of uncertainty and representativeness, i.e., the respective weights of uncertainty and representativeness; moreover these methods did not consider redundancy between selected samples. Some batch mode active learning methods will suffer from the same problem. In order to accelerate the learning process, it is necessary to speed up the learning process by selecting more than one sample at each iteration for batch mode active learning methods. So it needs to consider the diversity of the selected samples. To solve the above problems, this chapter proposes a multicriterion query-based active learning method. Experimental results show that our proposed method has achieved good performance.

2 Related Work

A general active learner can be modeled as a quintuple (G, Q, S, L, U) [8]. Initially, the training set L has few labeled samples to train the classifier G. After that, the query function Q is used to select a set of most informative samples from the unlabeled pool U and the supervisor S assigns a class label to each of them. Then, these new labeled samples are included into L and the classifier G is retrained using the updated training set. The closed loop of querying and retraining continues for some predefined iterations or until a stop criterion is satisfied.

There has been more research work in the sampling strategy of active learning. Uncertainty-based sampling methods are more commonly used. Entropy is most commonly used in uncertainty sampling method. Sample entropy can better represent the uncertainty of samples, i.e., the greater the entropy, the greater the uncertainty of the sample. However, in multiclass problems, the entropy does often not well reflect the uncertainty of the sample. Some may have larger classification uncertainty than the ones whose entropy may be higher. For the above problem, Joshi [9] proposed a more direct active learning sample selection criteria Best-versus-Second-Best (BvSB). The BvSB method considers the difference between the probability values of the two classes having the highest estimated probability value as a measure of uncertainty. The practical applications of the method get a better performance. Another common sampling strategy is based on the reduction of version space. Query-by-committee (QBC) algorithm is the most widely used famous algorithm. The committee, which is constituted by a set of group classifiers. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree. In essence, the QBC is based on uncertainty sampling. However, these methods only consider the impact on the labeled samples, without considering the distribution of unlabeled sample set, ignoring the relationship with unlabeled samples. The literature [7] shows that unlabeled samples have a great influence on classification accuracy. If the current sample can better represent the remaining unlabeled samples, then we say that the sample has a high-representative, meaning that the information content is higher. There have been some studies for a combination of uncertainty and representative aspects. Settles and Craven [7] proposed information density(ID) method, first with uncertainty methods measure the current sample basic information content, then use the sample feature vector cosine similarity method to calculate average similarity to all other samples in the input distribution, information is then multiplied by the density and set a fixed threshold to control the weight of density items. However, in many problems it is necessary to speed up the learning process by selecting more than one sample at each iteration. Li and Sethi [8], proposed that estimates the uncertainty level of each sample according to the output score of a classifier and selects only those samples whose output scores are within the uncertainty range. Patra and Bruzzone [10], proposed a fast clustering-based active learning method to solve the multiclass classification problems. However, these methods

select batch of samples at each iteration by considering only the uncertainty criterion. This will result in the selection of redundant samples which reduce the speed of the classifier without adding any additional information. To solve this problem, Brinker [11] presented a batch mode method of considering the diversity. In the literature [12, 13], the clustering method to measure the diversity was introduced into the design of active learning query function. For the combination of uncertainty, representation, and diversity, there is also some research. Lin and Bilmes [14] also studied batch mode active learning with submodular graph functions for the problem of training hidden Markov models for speech recognition, but this method is mainly designed for representativeness. Diversity, density, and relevance (analogous to uncertainty) are all incorporated in a query criterion by Xu [15] et al. but the approach is to simply interpolate three scores with two empirically-tuned weights. Tuning weights for active learning is more challenging in a real scenario than for classification accuracy.

Different from the above methods, we propose an active learning algorithm which integrates different selection criteria. The framework of our method consists of three key components: uncertainty, representativeness, and diversity criterion. Compared to the previous work this method has the advantages of (1) take into account the relationship with the labeled samples, but also make full use of the remaining unlabeled samples, which ensure the samples with higher uncertainty and better representation. In addition, considering the correlation between the selected samples, so that the selected samples are diverse; (2) without respecting to the weights of uncertainty, representativeness, and diversity. The first consideration of our method is the uncertainty of samples, and then with a combination of representativeness, and diversity. (3) Compared to the methods that directly deal with all the unlabeled samples, our method can greatly reduce the cost of sample selection, thereby improving efficiency.

3 MCQAL: Multicriterion Query-Based Active Learning

3.1 Problem Analysis

A limitation of the uncertainty sampling strategy is that it relies on the relative correctness or confidence of the current model, which can be a difficulty, especially in the early stages. And the uncertainty sampling can also suffer from a lack of exploration of the feature space and may not work well in some scenarios. As shown in Fig. 1, the red triangles and green circles represent the instances which have been labeled, the remaining white diamonds represent unlabeled sample set. Since sample x_A lies on the classification boundary, it must be selected by using uncertainty sampling strategy for human experts to label. In fact, x_B, x_C, x_D are samples with higher information content, they can better reflect the

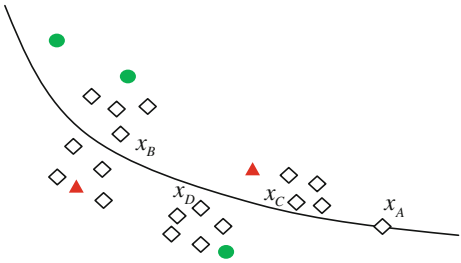

Fig. 1 An illustration of when uncertainty sampling can be a poor strategy

distribution of the sample set. Therefore, we make a combination of representativeness and diversity criteria based on the uncertainty sampling strategy, thus this part of samples will be selected together.

3.2 Uncertainty of Sample Selection

In the uncertainty sampling strategy, we employ the BvSB [9] approach, which consider the difference between the probability values of the two classes having the highest estimated probability value as a measure of uncertainty. Say that our estimated probability distribution for a certain example is denoted by P, where p_i denotes the membership probability for class i. Also suppose that the distribution P has a maximum value for class h. Based on current knowledge, the most likely set of classifiers in contention is C_h. The classification confidence for the classifiers in this set is indicated by the difference in the estimated class probability values, $p_h - p_i$. This difference is an indicator of how informative the particular example is to a certain classifier. Minimizing the difference $p_h - p_i$, or equivalently, maximizing the confusion (uncertainty), we obtain the BvSB measure.

$$\begin{aligned} \text{BvSB}^* &= \arg\min_{x_i \in U}\left(\min_{y \in Y, y \neq y_{\text{Best}}} \left(p(y_{\text{Best}}|x) - p(y|x)\right)\right) \\ &= \arg\min_{x_i \in U}\left(p(y_{\text{Best}}|x) - p(y_{\text{Second-Best}}|x)\right) \end{aligned} \tag{1}$$

3.3 Representativeness Measure

In addition to the most informative sample, we also prefer the most representative sample. The representativeness of a sample can be evaluated based on how many samples are similar or near to it. So, the samples with high-representative degree are less likely to be an outlier. Adding them to the training set will have an effect on a large number of unlabeled samples. In this section, we use the Gaussian

Process [16] model to measure the mutual information between the uncertain set and remaining unlabeled samples.

Using mutual information criterion, we define the mutual information-based representativeness measure for a candidate sample x_i as below

$$\mathrm{rep}(x_i) = I(x_i, U_{x_i}) = H(x_i) - H(x_i|U_{x_i}) \tag{2}$$

where U_{x_i} denotes the index set of unlabeled instances after removing x_i from U.

We propose to compute the entropy terms in (2) within a Gaussian Process framework. A Gaussian Process is a joint distribution over a (possibly infinite) set of random variables, such that the marginal distribution over any finite subset of variables is multivariate Gaussian. For our problem, we associate a random variable $\chi(x)$ with each instance. A symmetric positive definite Kernel function $K(\cdot,\cdot)$ is then used to produce the covariance matrix, such that

$$\sigma_i^2 = K(x_i, x_i) \tag{3}$$

$$\sum_{U_i U_i} = \begin{pmatrix} K(x_1,x_1) & K(x_1,x_2) & \dots & K(x_1,x_u) \\ K(x_2,x_1) & K(x_2,x_2) & \dots & K(x_2,x_u) \\ \vdots & \vdots & \vdots & \vdots \\ K(x_u,x_1) & K(x_u,x_2) & \dots & K(x_u,x_u) \end{pmatrix} \tag{4}$$

where the covariance matrix $\sum_{U_i U_i}$ is actually a kernel matrix defined over all the unlabeled instances indexed by U_i, we assume $U_i = \{1, 2, \dots, u\}$. One commonly used kernel function is the Gaussian kernel $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\lambda^2}\right)$.

Closed-form solutions exist for the entropy of multivariate Gaussian distributions such that

$$H(x_i) = \frac{1}{2}\ln\left(2\pi e \sum_{ii}\right) \tag{5}$$

$$H(x_i|U_{x_i}) = \frac{1}{2}\ln\left(2\pi e \sum_{i|U_i}\right) \tag{6}$$

using (5) and (6), the representativeness definition given in (2) can finally be rewritten into the following form

$$\mathrm{rep}(x_i) = H(x_i) - H(x_i|U_{x_i}) = \frac{1}{2}\ln\left(\frac{\sum_{ii}}{\sum_{i|U_i}}\right) \tag{7}$$

3.4 Diversity Analysis

Diversity criterion is to maximize the training utility of a batch. We prefer the batch in which the examples have high variance to each other. In this step we cluster all samples in the uncertainty set based on the representativeness measure proposed in Sect. 3.3. The samples in the same cluster may be considered similar to each other, so we will select the most informative sample from different clusters at one time. We apply the kernel k-means clustering algorithm.

In greater detail, let us assume that the kernel k-means clustering algorithm divides the m samples into h clusters $C_1, C_2, \ldots, C_h$. After $C_1, C_2, \ldots, C_h$ are obtained, the h most informative samples are selected as

$$x_k = \arg\min_{x \in C_k} \mathrm{rep}(x), k = 1, 2 \ldots, h. \tag{8}$$

3.5 A Combination Framework

In this section, we will study how to combine and strike a proper balance between these criteria, to reach the maximum effectiveness.

We first consider the uncertainty criterion. We choose m samples with the most informativeness score from all of the pool. By this preselecting, we make the selection process faster in the later steps since the size of uncertainty set is much smaller than that of the pool. Then we cluster the samples in uncertainty set and choose the centroid of each cluster into a batch. The centroid of a cluster is the most representative sample in that cluster since it has the largest density. Furthermore, the samples in different clusters may be considered diverse to each other. By this means, we consider representativeness and diversity criteria at the same time. We will summarize our overall algorithm, shown in Table 1.

4 Experimental Results

4.1 Design of Experiments

In order to assess the effectiveness of our algorithm, four international standard UCI data sets were used in the experiments data. Table 2 shows the information of data sets. As a comparison, we choose (1) information density (ID) active learning [7], a representative approach which selects informative and representative instances (2) A cluster-assumption-based batch mode active learning technique [17], a representative approach which selects informative and diverse instances.

Table 1 The pseudo-code of MCQAL is summarized in Algorithm 1

Algorithm 1:
Input: labeled data set L unlabeled data set U
Repeat
Training on L to get the probabilistic classification model Θ
for each x_i in U
Use (1) to measure the uncertainty of sample x_i
end for
Select the most uncertainty sample set MUSS;
for each x_j in MUSS
Use (7) to measure the representativeness of sample x_j
end for
Apply kernel k-means clustering algorithm to the MUSS;
Select one sample from each of the h clusters using (8)
Assign true labels to the h selected samples
$L = L \cup C^*$ $U = U \backslash C^*$
Until the stop criterion is satisfied

Table 2 Experimental data information table

Data set	Classes	Features	Unlabeled	Test
Ionosphere	2	34	246	105
Letters	26	16	10000	10000
Pen-Digits	10	16	7494	3498
Balance-scale	3	4	166	459

A RBF kernel with default parameters is used (performances with linear kernel are not as stable as that with RBF kernel). LibSVM is used to train a SVM classifier for all active learning approaches in comparison. To reduce the classification error, for every data set, we run the experiment for 10 times, each with a random partition of the data set. In the next section, we present the results of three different experiments. In the first experiment, we compared the accuracy of the proposed technique with other two techniques by using four datasets. The second experiment shows the diversity of samples during selecting. The computational load of the different methods is analyzed in the third experiment.

4.2 Results

Figure 2 shows the classification accuracy of different active learning approaches with varied numbers of queries.

For the two data sets Ionosphere and Letters, the proposed method is superior to the other two methods: When the number of labeled samples is small, our method

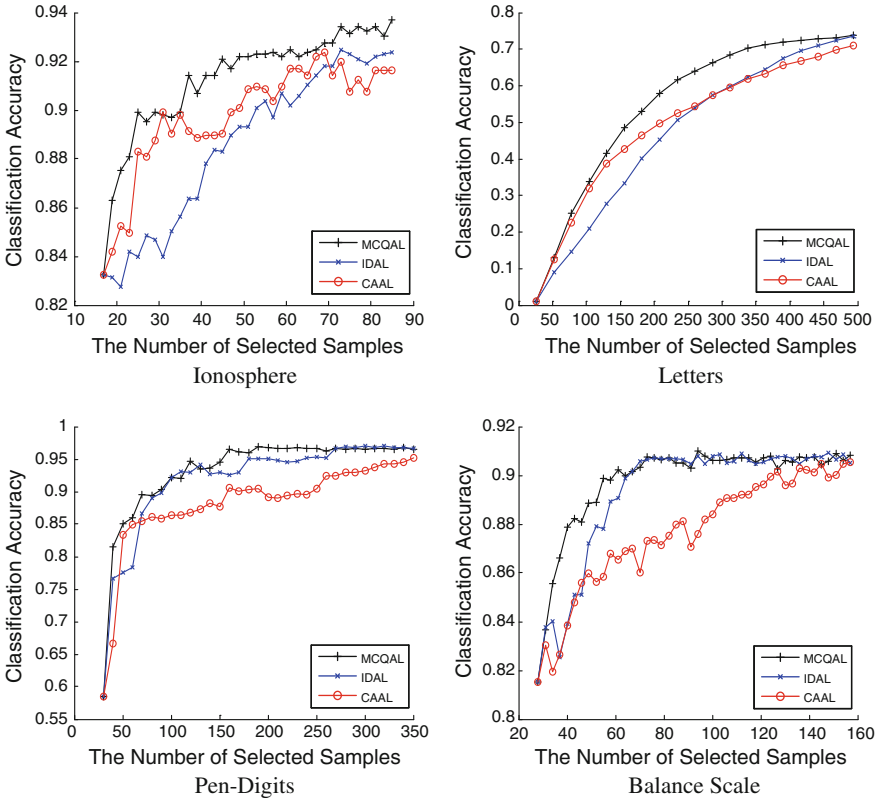

Fig. 2 Classification accuracy of different active learning approaches in four datasets

and CAAL are similar, but significantly better than IDAL; With the increase number of labeled samples, the superiority of our method gradually reflected and significantly better than CAAL. So in each iteration, the samples selected by our method are more informative (representative). In the condition of same labeled samples, our method is more effective to increase the classification accuracy.

Pen-Digits dataset, since our method took a priority on the basis of uncertainty and then selected samples both of representativeness and diversity, so the performance will be influenced by the uncertainty sample set. As shown in the section of the curve, we can see that the sample selected is not optimal. Those samples with high-representative but low-uncertainty may be ignored. In contrast, samples of higher information content are selected by IDAL.

The Balance Scale data set is imbalanced of class distribution. For imbalanced data sets, certain categories of samples are extremely rare, but often this part samples have higher information content. Therefore, it needs to pick up them as possible submitted to human experts for annotation. The result can be seen that our approach also received great performance on imbalanced dataset. Initially,

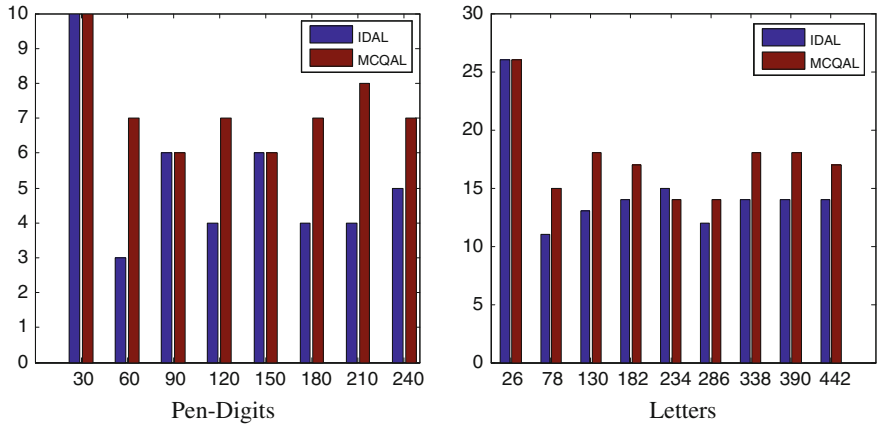

Fig. 3 The class number of selected samples at each iteration

Table 3 Computational time required by three techniques for all four data sets

Data set	Train-set	Test-set	Features	MCQAL (s)	CAAL (s)	IDAL (s)
Ionosphere	246	105	34	**2.17**	2.51	40.10
Letters	520	2080	16	**12.57**	19.57	44.19
Pen-digits	700	3498	16	**18.40**	73.58	200.86
Balance-scale	166	459	4	1.70	1.15	15.75

classification accuracy is much higher than the others as the selection of our approach is relatively more balanced. When the labeled samples reach a certain number, due to the high-valuable samples have been added to the training set, then the consideration of sample balance doesn't affect classification accuracy so obviously.

On the Pen-Digits and Letters datasets, each iteration we recorded the category number of selected samples. From the statistical data in the Fig. 3, in the condition of the same samples labeled at each iteration, the distribution of selected samples in our method is relatively more balanced. Reduce the selection redundancy while considering the diversity of samples, so the selected samples are more informative, quickly improving the classification performance.

The computational time required by the different techniques using the same experimental setting as described in the experiments. All the experiments were carried out on a PC (Pentium (R) Dual-Core CPU i5-2400@3.1 GHz, 4G RAM). Table 3 shows the computational time (in seconds) required by three techniques for all four data sets. From this table, compared with other two methods our method greatly reduced the running time in most cases.

5 Discussion and Conclusion

This paper presents a new active learning algorithm which combines uncertainty, representativeness, and diversity creation. On the basis of uncertainty sampling, we combined the measure of sample representativeness and analysis of sample diversity. This technique shortens the time required for training samples under the guarantee of classification accuracy. The labeling cost can be reduced without degrading the performance. For different data, our method needs to specify the size of the uncertain sample set according to the experiment. Therefore, according to the distribution of samples, how to dynamically determine the size of uncertainty set in order to ensure the optimal performance, is the focus of next study in the future.

Acknowledgments This work is partially supported by NSFC (No. 61003054, No. 61170020); College Natural Science Research project of Jiangsu Province (No. 10KJB520018); Science and Technology Support Program of Suzhou (No. SG201257); Science and Technology Support program of Jiangsu province (No. BE2012075); Open fund of Jiangsu Province Software Engineering R&D Center (SX201205). This work is also partially supported by the Natural Science Foundation of China under Grant No. 61003054 and 61170020, Jiangsu Province Colleges and Universities Natural Science Research Project under Grant No. 10KJB520018 and 13KJB520021, Jiangsu Province Science and Technology Support Program under Grant No. BE2012075, and Suzhou City Science and Technology Support Program under grant No. SG201257.

References

1. Demir B, Minello L, Bruzzone L (2013) An effective strategy to reduce the labeling cost in the definition of training sets by active learning. IEEE Geoscience and Remote Sensing Letters 11(1):79–83
2. Settles Burr (2012) Synthesis lectures on artificial intelligence and machine learning. Act Learn 6(1):1–114
3. Settles B (2009) Active learning literature survey. Computer science technical report 1648, University of Wisconsin-Madison, USA, pp 3–4
4. Settles B (2010) Active learning literature survey. University of Wisconsin-Madison, USA
5. Fu Y, Zhu X, Li B (2013) A survey on instance selection for active learning. Knowl Info Syst 35(2):249–283
6. Patra S, Bruzzone L (2012) A batch-mode active learning technique based on multiple uncertainty for SVM classifier. IEEE Geosci Remote Sens Lett 9(3):497–501
7. Settles B, Craven M (2008) An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the conference on empirical methods in natural language processing. association for computational linguistics, pp 1070–1079
8. Li M, Sethi IK (2006) Confidence-based active learning. Pattern Anal Mach Intell IEEE Trans 28(8):1251–1261
9. Joshi AJ, Porikli F, Papanikolopoulos N (2009) Multi-class active learning for image classification. Comput Vis Pattern Recogn 2372–2379
10. Patra S, Bruzzone L (2011) A fast cluster-assumption based active-learning technique for classification of remote sensing images. Geosci Remote Sens IEEE Trans 49(5):1617–1626

11. Brinker K (2003) Incorporating diversity in active learning with support vector machines. ICML 3:59–66
12. Liu R, Wang Y, Baba T et al (2008) SVM-based active feedback in image retrieval using clustering and unlabeled data. Pattern Recogn 41(8):2645–2655
13. Demir B, Persello C, Bruzzone L (2011) Batch-mode active-learning methods for the interactive classification of remote sensing images. Geosci Remote Sens IEEE Trans 49(3):1014–1031
14. Lin H, Bilmes J (2009) How to select a good training-data subset for transcription: submodular active selection for sequences. Washington University Seattle Department of Electrical Engineering, Washington
15. Xu Z, Akella R, Zhang Y (2007) Incorporating diversity and density in active learning for relevance feedback. Advances in Information Retrieval. Springer, Berlin, Heidelberg, pp 246–257
16. Kapoor A, Grauman K, Urtasun R et al. (2007) Active learning with gaussian processes for object categorization. IEEE 11th international conference on computer vision, pp 1–8
17. Patra S, Bruzzone L (2012) A cluster-assumption based batch mode active learning technique. Pattern Recogn Lett 33(9):1042–1048