Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 827586, 6 pages
http://dx.doi.org/10.1155/2014/827586

Research Article

An Active Learning Approach with Uncertainty, Representativeness, and Diversity

Tianxu He,[1] Shukui Zhang,[1,2] Jie Xin,[1,2] Pengpeng Zhao,[1] Jian Wu,[1] Xuefeng Xian,[1,3] Chunhua Li,[1] and Zhiming Cui[1,3]

[1] *School of Computer Science and Technology, Soochow University, Suzhou 215006, China*
[2] *State Key Lab. for Novel Software Technology, Nanjing University, Nanjing 210093, China*
[3] *Suzhou Vocational University, Suzhou 215104, China*

Correspondence should be addressed to Zhiming Cui; szzmcui@suda.edu.cn

Received 30 May 2014; Accepted 10 July 2014; Published 11 August 2014

Academic Editor: Juncheng Jia

Big data from the Internet of Things may create big challenge for data classification. Most active learning approaches select either uncertain or representative unlabeled instances to query their labels. Although several active learning algorithms have been proposed to combine the two criteria for query selection, they are usually ad hoc in finding unlabeled instances that are both informative and representative and fail to take the diversity of instances into account. We address this challenge by presenting a new active learning framework which considers uncertainty, representativeness, and diversity creation. The proposed approach provides a systematic way for measuring and combining the uncertainty, representativeness, and diversity of an instance. Firstly, use instances' uncertainty and representativeness to constitute the most informative set. Then, use the kernel k-means clustering algorithm to filter the redundant samples and the resulting samples are queried for labels. Extensive experimental results show that the proposed approach outperforms several state-of-the-art active learning approaches.

1. Introduction

According to an IDC report, the global data volume in 2014 has reached 8.7 ZB and will reach 40 ZB. With storage and transmission expanding PB level and EB level, it is indicated that big data will play an important role as important resources. Many supervised learning algorithms have been largely used in classification tasks [1, 2]. For a classification problem, the performance of classifier depends heavily on the labeled sample set. However, obtaining the labeled samples is very difficult while the labeled samples are scarce. In order to reduce the cost of labeling, active learning methods have been adopted to control the labeling process. Active learning is an effective method to solve these problems, which select high information content unlabeled samples to be labeled by experts [3, 4]. Querying the most informative instances is probably the most popular approach for active learning. Therefore, the querying strategy naturally becomes a research hotspot of active learning algorithms.

There are numerous different query strategies that have been used to decide which instances are most informative. The strategies are generally divided into two categories. One is based on uncertainty sampling [5, 6], which considers samples' uncertainty as information content and selects the most uncertain samples for labeling. Although most uncertainty query selection strategies have a wide range of applications and achieve good results in many circumstances, they fail to take information in the large amount of unlabeled instances into account and are prone to query outliers. Another category overcomes the disadvantages of uncertainty sampling and considers the samples' uncertainty and representativeness [7, 8].

In general, heuristic methods have been proposed to balance between the uncertainty and the representativeness of the selected sample. They encourage the selection of cluster centers. However, no measure has been taken to avoid repeating labeling samples in the same cluster. Namely, all methods above did not consider redundancy between

selected samples. Batch mode active learning methods will be affected by this problem. In order to accelerate the learning process, it is necessary to speed up the learning process by selecting more than one sample each iteration. So it needs to examine the diversity of the selected samples. To solve the above problems, we propose a novel active learning strategy that exploits information content measured by uncertainty, representativeness, and diversity of unlabeled instances. Samples selected for labeling are with high uncertainty and representativeness and little redundancy.

Our new query selection measure includes two steps. The first step is acquiring high information content samples set by combining uncertainty sampling and representativeness sampling. For the high informative samples set, we apply diversity sampling to get the final samples for labeling. The combination of the two terms is given in a general weighted product form and we use the kernel k-means clustering algorithm for diversity sampling. We conduct experiments on a few benchmark datasets and present promising results for the proposed active learning approach.

2. Related Work

A typical active learning framework assumes that there is a small set of labeled data L and a large pool U of unlabeled data available. Firstly, L is used to train the classifier C. Then, queries are selectively drawn from the pool, which is usually assumed to be closed. Typically, instances are queried in a greedy approach, according to an information measure used to estimate all instances in the pool, and labels for them are assigned by an expert. These new labeled samples are included into L and the classifier C is retrained. Querying loops continue for some predefined iterations or until a stop criterion is met.

A large number of active learning techniques have been introduced in the literature. Many methods employ an uncertainty sampling principle to select the unlabeled instance they are most hesitant to label. In [5], the most uncertain instance is taken as the one that has the largest entropy value on its probable labels. However, in multiclass problems, the entropy does not often 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 et al. [6] proposed a more effective active learning sample selection criterion BvSB. This criterion considers the difference between the probability values of the two classes having the highest estimated probability value as a measure of uncertainty, which results in a better performance in practical applications. Another common sampling strategy is based on the reduction of version space, among which query-by-committee (QBC) algorithm is the most popular one. QBC algorithms train a committee of classifiers and choose the instance on which the committee members most disagree [9]. In essence, the QBC is also based on uncertainty sampling. One immediate problem is that these approaches select samples close to the classification boundary in that they only consider uncertainty of samples, which are prone to be outliers. In order to avoid labeling outlier samples, representativeness sampling is an effective solution and there are some studies for a combination of uncertainty and representative aspects. References [10, 11] employ the unlabeled data by using the prior density as weights for uncertainty measures. A similar framework is proposed in [8], which uses a cosine distance to measure samples' representativeness. Literature [12] proposed an adaptive active learning method and showed better performance. Although different prediction models have been employed in these methods, they all ignore samples' cluster information or diversity information. Therefore these methods have the drawback of repeatedly labeling samples in the same cluster, which has little help for improving accuracy. In this paper, we develop a new active learning method which utilizes uncertainty sampling, representativeness sampling, and diversity sampling.

3. Proposed Approach

In this section, we present a novel active learning method that combines the three sampling criteria. The proposed active learning method has four key components: an uncertainty measure, a representativeness measure, an information content measure, and a diversity measure. We will introduce each of them below.

3.1. Uncertainty Measure. Uncertainty sampling aims to choose the most uncertain instance to label. We employ the best-versus-second-best (BvSB) [6] approach, which considers the difference between the probability values of the two classes having the highest estimated probability value as a measure of uncertainty. Assume that our estimated probability distribution for a certain example is denoted by P. Probability value of the best class guess and the second best guess are, respectively. We obtain the BvSB measure and refer [6] for detained information:

$$\text{Uncertainty}\left(x_i\right) = \text{BvSB} = \arg\min_{x_i \in U} \left(p\left(y_{\text{Best}} \mid x_i\right) - p\left(y_{\text{Second-Best}} \mid x_i\right)\right). \quad (1)$$

3.2. Representativeness Measure. As mentioned earlier, uncertainty sampling may suffer from the problem of selecting outlier samples. In order to prevent selecting these samples, representativeness sampling is an effective solution. The representativeness of a sample can be evaluated based on how many samples there are similar to it. So, samples with high representativeness are less likely to be outliers. In this section, we use the Gaussian Process [13] framework to measure the representativeness information between the current sample and the remaining unlabeled sample set.

Similar to the literature [12], we define representativeness measure for a candidate sample x_i as follows:

$$\text{Rep}\left(x_i\right) = H\left(x_i\right) - H\left(x_i \mid U_{x_i}\right), \quad (2)$$

where U_{x_i} denotes the set of unlabeled instances after removing x_i from U and $H(x_i)$ and $H(x_i \mid U_{x_i})$, respectively, represent entropies of x_i and the remaining unlabeled samples.

Input: labeled data set L unlabeled data set U
Repeat
  Training on L to get the probabilistic classification model C
  for each x_i in U
    Use (1) to measure the uncertainty of sample x_i
    Use (2) to measure the representativeness of sample x_i
    Use (3) to measure the information content of sample x_i
  end for
  Select the high information content set S;
  Apply kernel k-means clustering algorithm to S;
  Select k centers S_k from each of the clusters;
  Query true labels Y_k of the k selected samples;
  $L = L \cup \langle S_k, Y_k \rangle$
  $U = U \setminus \langle S_k, Y_k \rangle$
Output: final high-performance classifier C.

Algorithm 1: Incorporating uncertainty, representativeness, and diversity for active learning.

A Gaussian Process is a joint distribution over a set of random variables and the marginal distribution over any finite subset of variables is multivariate Gaussian. So we compute the entropy terms with it. For our issue, each instance is associated with a random variable. A symmetric kernel function $K(\cdot, \cdot)$ is then used to produce the covariance matrix, such that

$$\sum_{ii} = K\left(x_i, x_i\right), \tag{3}$$

$$\sum_{U_i U_i} = \begin{pmatrix} K\left(x_1, x_1\right) & K\left(x_1, x_2\right) & \cdots & K\left(x_1, x_n\right) \\ K\left(x_2, x_1\right) & K\left(x_2, x_2\right) & \cdots & K\left(x_2, x_n\right) \\ \vdots & \vdots & \vdots & \vdots \\ K\left(x_n, x_1\right) & K\left(x_n, x_2\right) & \cdots & K\left(x_n, x_n\right) \end{pmatrix}, \tag{4}$$

where the covariance matrix $\sum_{U_i U_i}$ is actually a kernel matrix defined over all the unlabeled instances and we assume $U_i = \{1, 2, \ldots, n\}$.

According to the property of multivariate Gaussian distribution, we can know that

$$\sum_{i \mid U_i}^{2} = \sum_{ii}^{2} - \sum_{iU_i} \sum_{U_i U_i}^{-1} \sum_{U_i i}. \tag{5}$$

Closed-form solutions exist for the entropy terms such that

$$\begin{aligned} H\left(x_i\right) &= \frac{1}{2} \ln\left(2\pi e \sum_{ii}\right), \\ H\left(x_i \mid U_{x_i}\right) &= \frac{1}{2} \ln\left(2\pi e \sum_{i \mid U_i}\right). \end{aligned} \tag{6}$$

Using (6), the representativeness definition can finally be in the following form:

$$\mathrm{Rep}\left(x_i\right) = \frac{1}{2} \ln\left(\frac{\sum_{ii}}{\sum_{i \mid U_i}}\right). \tag{7}$$

3.3. Information Content Measure. Given the uncertainty measure and the representativeness measure defined above, we seek to combine the strengths of both. The main idea is to pick samples that are not only with high uncertainty but also with high representativeness. We use the combination value of the two measures as information content value. The higher the combination value is, the higher the information content of corresponding sample is.

Specifically, we propose to combine the two values in a general product form and the information content of sample x_i is as follows:

$$\mathrm{Infor}\left(x_i\right) = \alpha * \mathrm{Uncertainty}\ \left(x_i\right) * \mathrm{Rep}\left(x_i\right), \tag{8}$$

where α is a tradeoff controlling parameter over the two terms. Samples with high $\mathrm{Infor}(x_i)$ value are more likely to be selected for labeling.

3.4. Diversity Measure. As we know, the high information content set may contain samples in the same cluster. In order to avoid selecting superfluous samples, we apply the kernel k-means clustering algorithm to cluster samples with high information content. We get the k clusters $C_1, C_2, \ldots, C_k$.

Then we choose the k cluster centers $S_k = \{x_{C_1}, x_{C_2}, \ldots, x_{C_k}\}$ for labeling, which can effectively guarantee that samples for labeling are with high information content and little redundancy.

First we consider representativeness and diversity criteria at the same time and get the high information content set S. Furthermore, diversity sampling is considered and redundant samples are filtered. So we cluster the samples in the high information content set and choose the clustering center of each cluster into a batch for labeling.

The overall framework of our active learning algorithm is given in Algorithm 1.

Table 1: Dataset properties and the corresponding sizes used.

Dataset	Classes	Features	Initial set size	Unlabeled set size	Test set size
USPS	10	256	30	5000	2000
Letters	26	16	30	5000	3000
Pendigits	10	16	30	7000	3498

Table 2: Cluster number for each dataset.

Dataset	Labeled numbers at each round
USPS	10
Pendigits	10
Letters	26

4. Experimental Results

In order to evaluate the effectiveness of our proposed approach described in previous sections, we demonstrate results on three UCI datasets: the Letter Recognition Data Set, USPS: optical recognition of handwritten digits originally from the US Postal Service, and Pendigits: pen-based recognition of handwritten digits. The chosen datasets and their properties are summarized in Table 1 along with initial samples set, unlabeled samples set, and test set sizes used in our experiments.

The experiments are conducted to compare the proposed active learning approach to a number of active learning methods, including (1) BvSB [6], which is the uncertainty sampling method, and (2) information density (ID), which denotes the active learning method in [8] that uses the cosine distance to measure an information density and selects uncertain and representative instances.

LibSVM is employed to train the train a SVM classifier for all these approaches, and it provides probabilistic predictions over the class labels.

4.1. Size of k for Each Dataset. Table 2 shows the number of cluster for each dataset. Note that class numbers of each dataset have already been known in advance. Then, we adjust the k parameter according to the class number of each dataset to cluster current high information content samples set.

4.2. Classification Accuracy on Three Datasets. In Figure 1, we show results on the USPS dataset, a dataset consisting of handwritten digits from the US Postal Service. At each active learning round, we select 10 samples for labeling. At early iterations, performances of all methods are similar. As the number of labeled samples increases, our method gradually dominates the other two approaches and the proposed approach selects the most useful samples.

This difference between the three active selection methods becomes more clearly when we look at the results on the Letters dataset. Similar to USPS dataset, we select 10 samples for labeling. We can know that, for achieving the same value of classification accuracy on the test data, our method needs far fewer training samples than the other two approaches. Note that ID method does marginally better than most uncertain sampling. The difference can be attributed to the fact that ID method combines uncertainty and representativeness of samples while most uncertain sampling only considers uncertainty of samples.

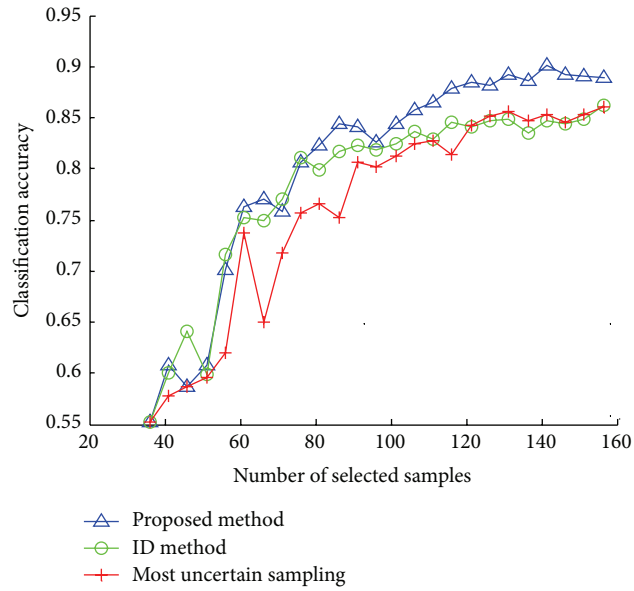

Figure 1: Results on USPS dataset.

Figure 3 shows classification accuracy plots on the Letters dataset, which has 26 classes. Most uncertain sampling and ID method perform even worse on this problem due to the larger number of classes. They give a bad indicator of information content of unlabeled samples in this case, and they give comparable poor performance. Even with a larger number of classes, the figure indicates that our approach outperforms other active selection methods.

4.3. Comparison of Diversity. All the results show that our approach in selecting diversity samples is very effective, especially in Pendigits dataset (see Figure 2). From Figures 4 and 5, our method selects samples included in all classes while the other two methods only choose samples with uncertainty or representativeness which is distributed in only a part of classes. Therefore, the proposed method performs better than other methods in most cases.

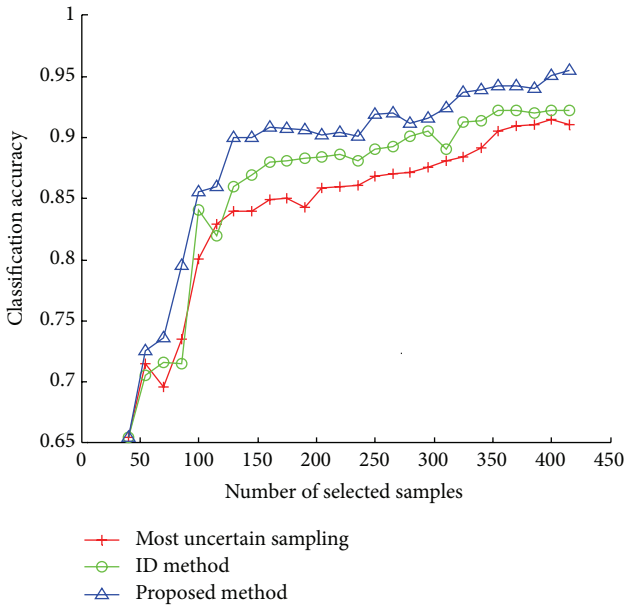

Figure 2: Results on Pendigits dataset.

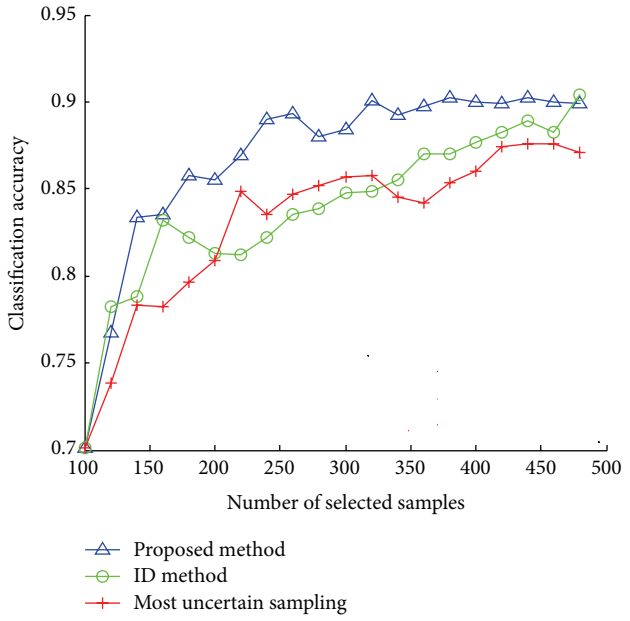

Figure 3: Results on Letters dataset.

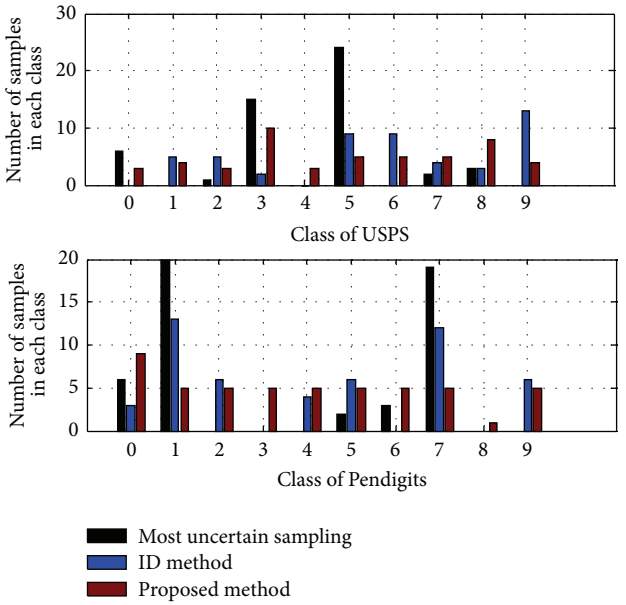

Figure 4: Comparison of diversity on 10 classes.

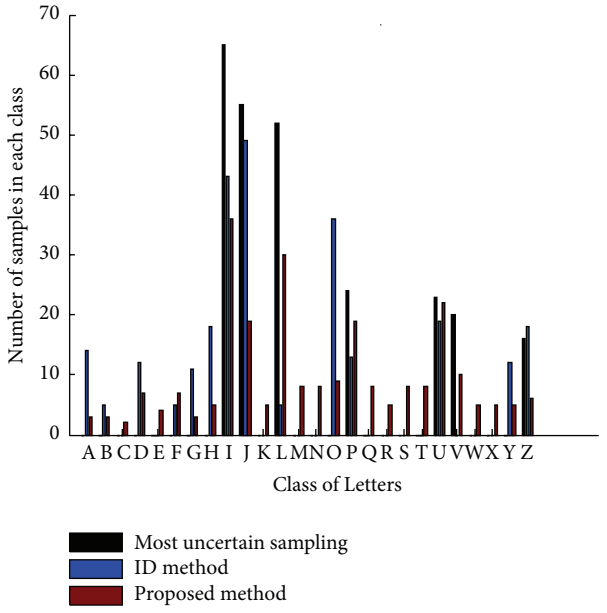

Figure 5: Comparison of diversity on 26 classes.

5. Conclusion and Future Work

In this paper, we presented a novel adaptive active learning approach which combines uncertainty measure and representativeness measure with diversity measure together to conduct samples selection. The proposed method can select samples with high information content and little redundancy. Experiments on multiple datasets show advantages of our approach. The expert in our approach is assumed to be accurate, indefatigable (always answers the queries), and insensitive to costs. Labeling an optimal utility subset is still costly and expensive in many cases. Crowdsourcing labelers, which are composed of some cheap and noisy labelers, have now been considered for active learning. Future work will extend to reduce the cost of this issue.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the Natural Science Foundation of Jiangsu Province under Grant no. BK2011376, The Opening Project of Suzhou High-Tech Key Laboratory of Cloud Computing & Intelligent Information Processing no. SXZ201302, Science and Technology Support Program of Suzhou no. SG201257, Science and Technology Support Program of Jiangsu Province no. BE2012075, Suzhou Key Laboratory of Converged Communication (no. SKLCC2013XX), The Opening Project of Suzhou High-Tech Key Laboratory of Cloud Computing & Intelligent Information Processing no. SXZ201302, Provincial Key Laboratory for Computer Information Processing Technology (no. KJS1329), Open Fund of Jiangsu Province Software Engineering R&D Center no. SX201205, and Jiangsu College Graduate Research and Innovation Plan no. CXLX12_0810. This work is also supported by the National Natural Science Foundation of China under Grants nos. 61070169, 61003054, and 61170020 and Jiangsu Province Colleges and Universities Natural Science Research Project under Grant no. 13KJB520021.

References

[1] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601–614, 2012.

[2] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[3] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, pp. 55–66, 2010.

[4] Y. Fu, X. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowledge and Information Systems*, vol. 35, no. 2, pp. 249–283, 2013.

[5] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[6] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2372–2379, June 2009.

[7] Z. Xu, R. Akella, and Y. Zhang, "Incorporating diversity and density in active learning for relevance feedback," in *Advances in Information Retrieval*, pp. 246–257, Springer, Berlin, Germany, 2007.

[8] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pp. 1070–1079, Association for Computational Linguistics, October 2008.

[9] S. Majidi and G. Crane, "Committee-based active learning for dependency parsing," in *Research and Advanced Technology for Digital Libraries*, vol. 8092 of *Lecture Notes in Computer Science*, pp. 442–445, Springer, Berlin, Germany, 2013.

[10] C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 260–268, 2002.

[11] A. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proceedings of the 15th International Conference on Machine Learning*, 1998.

[12] X. Li and Y. Guo, "Adaptive active learning for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 859–866, Portland, Ore, USA, June 2013.

[13] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with gaussian processes for object categorization," in *Proceeding of the 11th IEEE International Conference on Computer Vision (ICCV '07)*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.