# Adaptive Active Learning for Image Classification

Xin Li Yuhong Guo
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122
{xinli,yuhong}@temple.edu

## Abstract

*Recently active learning has attracted a lot of attention in computer vision field, as it is time and cost consuming to prepare a good set of labeled images for vision data analysis. Most existing active learning approaches employed in computer vision adopt most uncertainty measures as instance selection criteria. Although most uncertainty query selection strategies are very effective in many circumstances, they fail to take information in the large amount of unlabeled instances into account and are prone to querying outliers. In this paper, we present a novel adaptive active learning approach that combines an information density measure and a most uncertainty measure together to select critical instances to label for image classifications. Our experiments on two essential tasks of computer vision, object recognition and scene recognition, demonstrate the efficacy of the proposed approach.*

## 1. Introduction

Image classification has a long history in computer vision research, and it remains a major challenge due to the broad intra-class diversity of images caused by shape, color, size, or environmental conditions. To build a robust image classifier, it typically requires a large number of labeled training instances. For example, 10,000 instances of handwriting digits are used for training classifiers in [33]. It is time and cost consuming to prepare such a large set of labeled training instances. On the other hand, one fascinating characteristic of human vision system is that we can categorize image objects with only few labeled training instances. Is it possible for a computer to achieve this with the solid support of machine learning techniques? This is the motivation of this research. We aim to develop an effective active learning method to build a competitive classifier with a limited amount of labeled training instances.

Training a good classifier with minimal labeling cost is a critical challenge posed in machine learning research. Randomly selecting unlabeled instances to label is inefficient in many situations, since non-informative or redundant instances might be selected. Aiming to reduce labeling effort, active learning methods have been adopted to control the labeling process. Recently, active learning has been studied in computer vision [3, 14, 13, 15, 16], focusing on pool-based setting. These works however merely evaluate the informativeness of instances with most uncertainty measures, which assume an instance with higher classification uncertainty is more critical to label. Although the most uncertainty measures are effective on selecting informative instances in many scenarios, they only capture the relationship of the candidate instance with the current classification model and fail to take the data distribution information contained in the unlabeled data into account. This may lead to selecting non-useful instances to label. For example, an outlier can be most uncertain to classify, but useless to label. This suggests representativeness of the candidate instance in addition to the classification uncertainty should be considered in developing an active learning strategy.

In this paper, we propose a novel adaptive active learning strategy that exploits information provided by both the labeled instances and the unlabeled instances for query selection. Our new query selection measure is an adaptive combination of two terms: an uncertainty term based on the current classifier trained on the labeled instances; and an information density term that measures the mutual information between the candidate instance and the remaining unlabeled instances. The combination of the two terms is given in a general weighted product form. We seek to obtain an adaptive combination of the two terms by selecting the weight parameter to minimize the expected classification error on unlabeled instances. We conduct experiments on a few benchmark image classification datasets and present promising results for the proposed active learning method.

## 2. Related Work

A large number of active learning techniques have been developed in the literature. Most of them have been focused on selecting a single most informative unlabeled instance to label each time. Many such approaches make myopic decisions based solely on the current learned classifiers and employ an *uncertainty sampling* principle to select the unlabeled instance they are most uncertain to label. In [18, 26], the most uncertain instance is taken as the one that has the largest entropy on the conditional distribution over its labels. Support vector machine methods choose the most uncertain instance as the one that is closest to the classification boundary [2, 25, 28]. *Query-by-committee* algorithms train a committee of classifiers and choose the instance on which the committee members most disagree [9, 19].

One apparent shortcoming of the active learning strategies reviewed above is that they select a query based only on how that instance relates to the current classifier(s), whereas ignoring the large set of unlabeled instances. One immediate problem is that these approaches are prone to querying outliers, as we discussed before. Moreover, the goal of active learning is producing a classifier that has good generalization performance on unseen instances in the problem domain. Although it might not be possible to access the domain distribution directly, relevant information can be obtained from the large pool of unlabeled instances. Many active learning methods have been proposed to exploit unlabeled data to minimize the generalization error of the trained classifier. In [24], queries are selected to minimize the generalization error in a direct way by maximizing the *expected error reduction* on unlabeled data with respect to the estimated posterior label probabilities. Another class of active learning approaches minimize the generalization error indirectly by *reducing model variances*, including a statistical approach [4], and a similar approach that selects optimal queries based on Fisher information [35]. These generalization error minimization approaches are generally *computationally expensive*. An alternative class of active learning methods use a number of heuristic measures to exploit the information in unlabeled data. The methods in [19, 32] employ the unlabeled data by using the prior density $p(\mathbf{x})$ as weights for uncertainty measures. A similar framework is employed in [26], which uses a cosine distance to measure an information density. The methods in [6, 20] explicitly combine clustering and active learning together to exploit both labeled and unlabeled instances. In [10, 17], instances are selected to maximize the *increase of mutual information* between the selected set of instances and the remaining ones based on Gaussian Process models. The method in [23] extends the query-by-committee algorithm by exploiting unlabeled data. The work [11] seeks the instance whose optimistic label provides *maximum mutual information* about the labels of the remaining unlabeled instances, which implicitly exploits the clustering information contained in the unlabeled data in an optimistic way.

In the realm of computer vision, researchers have adopted active learning in image/video annotation [16, 34, 31], image/video retrieval [29, 12] and image/video recognition [30, 15, 13, 22, 14]. The work [29] applies active learning on object detection and the approach aims to deal with a large amount of images crawled online. The work [14] generalizes the margin-based uncertainty measure to the multi-class case. In [22], a two dimensional active learning method is proposed to conduct selection over instance-label pairs instead of solely instances. The work [13] introduces a probabilistic variant of a KNN method used for active learning. The work [15] uses Gaussian Process as a probabilistic prediction model to gain a direct estimate of uncertainty measure for active learning in binary classification case. Although different prediction models have been employed in these methods, they all used the simple *uncertainty sampling* active learning strategy for instance selection. Therefore these methods have the drawback of ignoring the distributional information contained in the large number of unlabeled instances, as we discussed above. In this paper, we develop a new active learning method for image classification tasks, which overcomes the inherent limitation of uncertainty sampling.

## 3. Proposed Approach

Different active learning strategies have different strengths in identifying which instance to query given current classifier. In this section, we present a novel active learning method that combines the strengths of different active learning strategies in an adaptive way. The proposed active learning method has three key components: an uncertainty measure, an information density measure and an adaptive combination framework. We will introduce each of them below. Moreover, our approach is based on probabilistic classification models. We use logistic regression as our probabilistic classification model in the experiments.

**Notations.** We use the following notations in this paper. We use $\mathbf{x}_i \in \mathbb{R}^d$ to denote the input feature vector of the $i$th instance, and $y_i \in \{1, \cdots, K\}$ to denote its class label. We use $L$ and $U$ to denote the index sets of the labeled and unlabeled instances respectively. Assuming we are initially given a set of labeled instances $\{(\mathbf{x}_i, y_i)\}_{i\in L}$ and a large set of unlabeled instances $\{\mathbf{x}_i\}_{i\in U}$, we aim to sequentially select the most informative instances from $U$ to query and move them into the labeled set $L$ such that a good classifier can be trained on instances indexed by $L$.

### 3.1. Uncertainty Measure

Uncertainty sampling is one simplest and most commonly used active active learning strategy. It aims to choose the most uncertain instance to label. For probabilistic classification models, the uncertainty measure is defined as the conditional entropy of the label variable $Y$ given the candidate instance $\mathbf{x}_i$:

$$f(\mathbf{x}_i) = H(Y|\mathbf{x}_i, \theta_L) \tag{1}$$
$$= -\sum_{y \in \mathcal{Y}} P(y|\mathbf{x}_i, \theta_L) \log P(y|\mathbf{x}_i, \theta_L)$$

where $\mathcal{Y}$ denotes the set of all class values, $\theta_L$ represents the classification model trained over the labeled set $L$, and the conditional distribution $P(y|\mathbf{x}_i, \theta_L)$ is determined using this model. This uncertainty measure captures the informativeness of the candidate instance with respect to the labeled instances. The uncertainty sampling active learning strategy selects the candidate unlabeled instance $\mathbf{x}_{i^*}$ that has the largest conditional entropy

$$i^* = \arg\max_{i \in U} H(Y|\mathbf{x}_i, \theta_L) \tag{2}$$

As we mentioned before, the uncertainty sampling approaches are limited in that their assessment of an instance involves only the small set of currently labeled instances (that produce the classifier $\theta_L$) but not the distribution of the other unlabeled instances.

### 3.2. Information Density Measure

To cope with the drawback of uncertainty sampling, we next take the unlabeled instances into consideration when selecting an instance to query. Our motivation is that the representative instances of the input distribution can be very informative for improving the generalization performance of the target classifier. Although the input distribution is usually not given, we have a large set of unlabeled instances that can be used to approximate the input space. It has been shown in previous semi-supervised learning work that the distribution of unlabeled data is very useful for training good classification models [5, 27]. Intuitively, one would prefer to select the instance that is located in a *dense region* regarding the other unlabeled instances, since such an instance will be much more *informative* about other unlabeled instances than the ones located in a sparse region. We thus use the term *information density* to indicate the informativeness of a candidate instance for the remaining unlabeled instances. Specifically, in this work, we define the information density measure as the mutual information between the candidate instance and the remaining unlabeled instances within a Gaussian Process framework.

Mutual information is a quantity that measures the mutual dependence of two sets of variables, which is a more straightforward representativeness measure than the marginal density $p(\mathbf{x})$ used in [19, 32, 27], and a more principled representativeness measure than the cosine distance information density measure used in [26]. We define the mutual information based information density measure for a candidate instance $\mathbf{x}_i$ as below

$$d(\mathbf{x}_i) = I(\mathbf{x}_i, \mathbf{X}_{U_i}) = H(\mathbf{x}_i) - H(\mathbf{x}_i|\mathbf{X}_{U_i}) \tag{3}$$

where $U_i$ denotes the index set of unlabeled instances after removing $i$ from $U$, such that $U_i = U - i$, and $\mathbf{X}_{U_i}$ denotes the set of instances indexed by $U_i$.

We propose to compute the entropy terms in (3) 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 $\mathcal{X}(\mathbf{x})$ with each instance $\mathbf{x}$. A symmetric positive definite Kernel function $\mathcal{K}(\cdot, \cdot)$ is then used to produce the covariance matrix, such that $\sigma_i^2 = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i)$ and

$$\Sigma_{U_iU_i} = \begin{pmatrix} \mathcal{K}(\mathbf{x}_1, \mathbf{x}_1) & \mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathcal{K}(\mathbf{x}_1, \mathbf{x}_m) \\ \mathcal{K}(\mathbf{x}_2, \mathbf{x}_1) & \mathcal{K}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathcal{K}(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{K}(\mathbf{x}_m, \mathbf{x}_1) & \mathcal{K}(\mathbf{x}_m, \mathbf{x}_2) & \cdots & \mathcal{K}(\mathbf{x}_m, \mathbf{x}_m) \end{pmatrix}$$

where we assume $U_i = \{1, \cdots, m\}$. Thus the covariance matrix $\Sigma_{U_iU_i}$ is actually a kernel matrix defined over all the unlabeled instances indexed by $U_i$. One commonly used kernel function is the Gaussian kernel

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{(\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\tau^2}\right). \tag{4}$$

According to the property of multivariate Gaussian distribution, the conditional distribution of $P(\mathbf{x}_i|\mathbf{X}_{U_i}) = P(\mathcal{X}(\mathbf{x}_i)|\mathcal{X}(\mathbf{X}_{U_i}))$ is also a Gaussian distribution with a conditional covariance matrix

$$\sigma^2_{i|U_i} = \sigma_i^2 - \Sigma_{iU_i}\Sigma^{-1}_{U_iU_i}\Sigma_{U_ii} \tag{5}$$

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

$$H(\mathbf{x}_i) = \frac{1}{2}\ln\left(2\pi e\sigma_i^2\right) \tag{6}$$
$$H(\mathbf{x}_i|\mathbf{X}_{U_i}) = \frac{1}{2}\ln\left(2\pi e\sigma^2_{i|U_i}\right) \tag{7}$$

Using (6) and (7), the information density definition given in (3) can finally be rewritten into the following form

$$d(\mathbf{x}_i) = \frac{1}{2}\ln\left(\frac{\sigma_i^2}{\sigma^2_{i|U_i}}\right) \tag{8}$$

which is easily computable without using any classification model given a positive definite kernel function.

### 3.3. A Combination Framework

Given the uncertainty measure and the information density measure defined above, we aim to develop a combination framework to integrate the strengths of both. The main idea is to pick the instance that is not only most uncertain to classify based on the current classifier, but also very informative about the remaining unlabeled instances. Thus after adding this instance to the labeled set, the new classifier produced can make more accurate predictions on the unlabeled instances. Specifically, we propose to combine the two measures in a general product form of combination framework as below

$$h_\beta(\mathbf{x}_i) = f(\mathbf{x}_i)^\beta d(\mathbf{x}_i)^{1-\beta} \tag{9}$$

where $0 \leq \beta \leq 1$ is a tradeoff controlling parameter over the two terms. For the combination measure given in Eq. (9), although the uncertainty term $f(\mathbf{x}_i)^\beta$ is a discriminative measure, the information density term $d(\mathbf{x}_i)^{1-\beta}$ is computed in the input space and has no direct connection with the target discriminative classification model. Using such a heuristic combination measure, we aim to pick the most informative instance for reducing the generalization error of the classification model without the computationally expensive steps of retraining classification model for each candidate instance.

The only computationally expensive operation for this information density assisted combination measure is the matrix inversion operation $\Sigma^{-1}_{U_iU_i}$ used to compute the conditional covariance $\sigma^2_{i|U_i}$ in Eq. (5). It is very inefficient to compute a matrix inverse $\Sigma^{-1}_{U_iU_i}$ for each candidate instance $i \in U$. We tackle this computational issue by borrowing a fast algorithm from [36] to compute the inverse matrix with one row/column removed. The basic idea is that for any $i \in U$, we can compute the inverse matrix $\Sigma^{-1}_{U_iU_i}$ from the given $\Sigma_{UU}$ and $\Sigma^{-1}_{UU}$ directly without matrix inversion; see [36] for details. Thus we only need to conduct one matrix inversion at the beginning of the active learning process. Moreover, one can use subsampling to further reduce the computational cost for large unlabeled sets. That is, in each iteration of active learning, one can first randomly sample a subset of unlabeled instances, and then restrain the candidate instance selection to this subset.

A similar combination strategy to our proposed one in Eq. (9) has been presented in [26] in form of $[f(\mathbf{x}_i)d(\mathbf{x}_i)^\beta]$. However, it uses the average cosine distance between the candidate instance and all unlabeled instances as its information density measure. Moreover, it uses a predefined weight parameter $\beta$. Below we propose to adaptively select the best $\beta$ from a range of pre-defined values to use in each iteration of active learning.

### 3.4. Adaptive Combination

One important issue regarding the combination strategy we proposed above is to select a proper weight parameter $\beta$ for $0 \leq \beta \leq 1$. The $\beta$ value controls the degree of relative importance of the two component measures. When $\beta > 0.5$, the uncertainty measure is treated as a more important measure than the information density measure since more weights are put on the uncertainty measure. In the extreme case of $\beta = 1$, it is equivalent to most uncertainty sampling. Similarly, when $\beta < 0.5$, more weights are put on the information density measure. However, it is difficult to pre-define the relative importance of the two measures for each different dataset. Moreover, the relative importance of the two measures can be dynamically changing across different iterations and stages of the active learning process. To achieve the best possible instance selection in each iteration, one thus needs to dynamically evaluate the relative informativeness of the two measures and determine the $\beta$ value for each instance selection. Unfortunately this is a very difficult problem to solve.

In this work, we propose to take a simple nonmyopic step to adaptively pick the $\beta$ value from a set of pre-defined candidate values. Specifically, in each iteration of active learning, we compute the uncertainty measure $f(\mathbf{x}_i)$ and the information density measure $d(\mathbf{x}_i)$ for each candidate instance $\mathbf{x}_i$. Then we select a set of $b$ instances using $b$ different $\beta$ values from a pre-defined set $B$ according to the combination measure $h_\beta(\mathbf{x}_i)$ defined in Eq. (9). For example, for a given set $B = [0.1, 0.2, \ldots, 0.9, 1]$, we can select $b = 10$ instances, one for each different $\beta$ value from this set. Then selecting the best $\beta$ value is equivalent to selecting the most informative instance from the $b$ selected instances. We propose to make this selection by minimizing the expected classification error on the unlabeled instances. Let $S$ denote the set of $b$ selected instances. For each candidate instance $\mathbf{x}$ from the set $S$, we label it with a label value $y$ with probability $P(y|\mathbf{x}, \theta_L)$. By adding each possible instance-label pair $< \mathbf{x}, y >$ into the current labeled set $L$ and retraining the classifier on the augmented labeled set, we can measure the prediction loss of the new classifier on all unlabeled instances. The expected loss of the candidate instance $\mathbf{x}$ can be computed as a weighted sum of the prediction loss obtained using all possible labels $y$ under the distribution $P(y|\mathbf{x}, \theta_L)$. Specifically, we conduct instance selection from the set $S$ using the following equation

$$\mathbf{x}^* = \tag{10}$$
$$\arg\min_{\mathbf{x}\in S} \sum_{y\in\mathcal{Y}} P(y|\mathbf{x}, \theta_L)\Big(\sum_{i\in U}(1 - P(\widehat{y}_i|\mathbf{x}_i, \theta_{L+<\mathbf{x},y>}))\Big)$$

where $\theta_{L+<\mathbf{x},y>}$ denotes the new model parameter after retraining on the augmented set $L+ < \mathbf{x}, y >$, and $\widehat{y}_i$ is the predicted label for instance $\mathbf{x}_i$.

The overall active learning algorithm is given in Algorithm 1. Although classifier retraining is required to compute the expected classification loss, this only needs to be done for a very small number of pre-selected instances in $S$. The computational cost can thus be maintained within a reasonable range.

**Algorithm 1** Adaptive Active Learning Algorithm

```
input: Labeled set $L$, Unlabeled set $U$,
       and $B = [0.1, 0.2, \ldots, 1]$
repeat
  Training a probabilistic classifier $\theta_L$ on $L$.
  for $i \in U$ do
    Compute $f(\mathbf{x}_i)$ using Eq.(1).
    Compute $d(\mathbf{x}_i)$ using Eq.(8).
    Compute $h_\beta(\mathbf{x}_i)$ with different $\beta \in B$ via Eq.(9).
  end for
  Let $S = \emptyset$.
  for $\beta \in B$ do
    Select an instance $\mathbf{x} = \arg\max_{i \in U} h_\beta(\mathbf{x}_i)$.
    Put $\mathbf{x}$ into set $S$, $S = S \cup \mathbf{x}$.
  end for
  Select instance $\mathbf{x}^*$ from $S$ using Eq. (10).
  Remove $\mathbf{x}^*$ from the unlabeled set $U$.
  Query the true label $y^*$ of $\mathbf{x}^*$, and update $L$ by
        adding $< \mathbf{x}^*, y^* >$ into it.
until enough instances are queried
```

# 4. Experimental Results

We evaluated the effectiveness of the proposed approach on three image classification datasets, one dataset for scene recognition and two datasets for object recognition. We report our experimental results in this section.

## 4.1. Experimental Setting

**Datasets** For scene recognition task, we used the *13 Natural Scene Categories* dataset [8] (a superset of MIT Urban and Natural Scene dataset [21]), which consists of both natural (coast, forest, mountain, etc.) and man-made scenes (kitchen, tall building, street, etc.), and has 3859 images in total. For object detection tasks, we used the *Caltech101* and *Pascal VOC 2007* datasets. *Caltech101* [7] is a benchmark dataset for object recognition, which contains 102 categories (including the *background* category), and has 8677 images in total. Instead of using the entire set, we randomly selected 30 images from each category to form a subset of Caltech101, which has 3060 images in total. *Pascal VOC 2007* is a widely used dataset for object recognition in computer vision community. We used its training and validation data which contains 5011 images. Since we are not solving multi-label problems, the images with more than one labels are simply discarded. The dataset we finally obtained contains 2989 images from 20 object categories.

**Approaches** The experiments are conducted to compare the proposed adaptive active learning approach to a number of active learning methods, including (1) *Random Sampling*; (2) *Most Uncertainty*, which is the uncertainty sampling method; (3) *Near Optimal*, which is the mutual information based active learning method proposed in [10]; and (4) *Fixed Combination*, which denotes the active learning method in [26] that uses the cosine distance to measure an information density $d(\mathbf{x})$ and uses a fixed $\beta$ parameter to produce a combination measure $[f(\mathbf{x})d(\mathbf{x})^\beta]$. We employed logistic regression as the classification model for all these approaches, and it provides probabilistic predictions over the class labels.

## 4.2. Experiment I: Scene Recognition

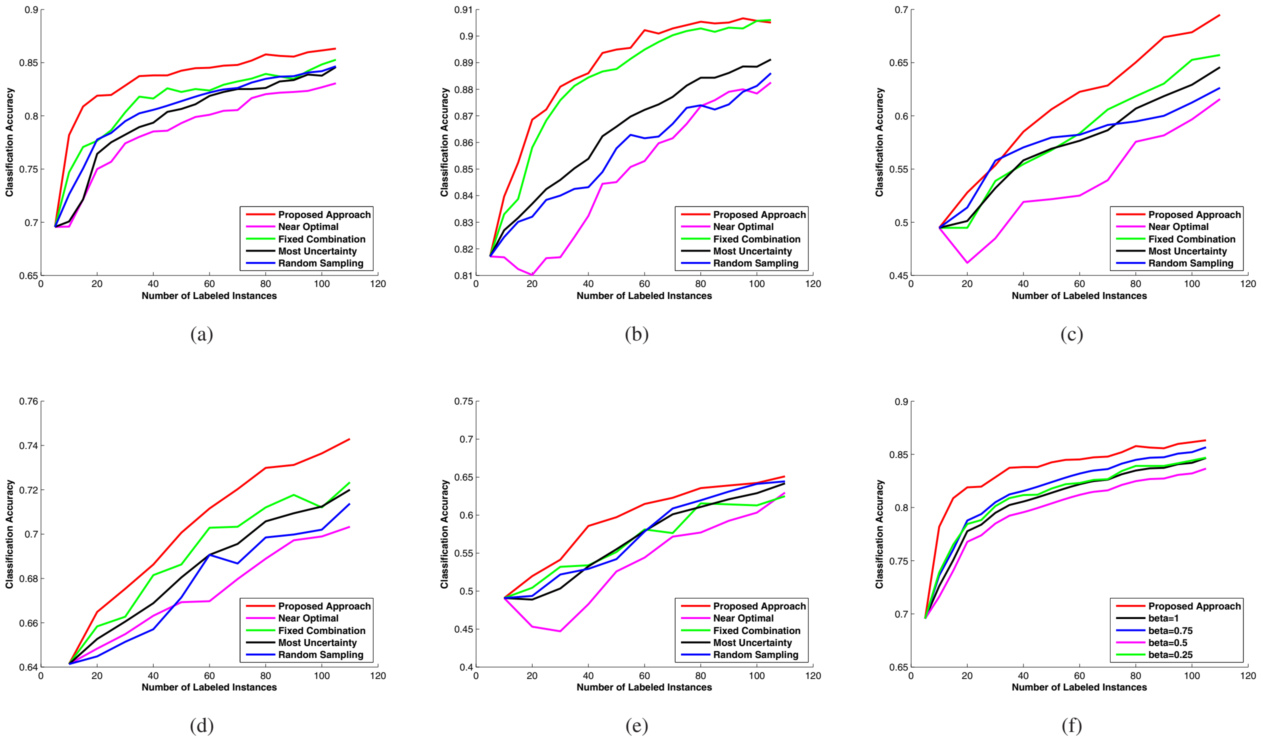
First, we conducted experiments on the 13 Natural Scene dataset using GIST features [21]. We randomly selected two subsets with 5 classes and three subsets with 10 classes from the 13 Natural Scene dataset. On each subset, the selected instances are randomly partitioned into labeled instances, unlabeled instances and testing instances according to a proportion of 2%, 68% and 30% respectively. Each active learning algorithm starts with the labeled instances and iteratively selects instances from the unlabeled set to label, one at each time, with maximum 100 iterations. After each selected instance being labeled, a logistic regression classifier is trained on the labeled data and tested on the test data to record its classification accuracy.

The experiments are repeated 10 times and the average results are reported in Figure 1. In Figure 1a, the classifier achieves high performance with much fewer iterations in our proposed active learning approach than in other approaches, which demonstrates that the proposed active learning strategy selects more effective queries. In Figure 1b, the advantages of the proposed approach over *Random Sampling*, *Most Uncertainty* Sampling and *Near Optimal* method are obvious, but the difference between it and *Fixed Combination* is small, which suggests the fixed $\beta$ parameter used in *Fixed Combination* happens to fit to this subset. The Figure 1c–Figure 1e show the results on three 10-class subsets of the 13 Natural Scene dataset. With the principled information density measure and the adaptive integration framework, the proposed adaptive active learning approach consistently outperforms all the other methods across these experiments regardless of the number of classes. We also compared the adaptive method with methods that use our proposed combination framework but with different fixed $\beta$ values, $\beta \in \{0.25, 0.5, 0.75, 1\}$, on the 5-class subset of Figure 1a. The results are reported in Figure 1f, which illustrates the effectiveness of selecting adaptive $\beta$ values against using fixed $\beta$ values.

## 4.3. Experiment II: Object Recognition

The second set of experiments are conducted on the two object recognition datasets, *Pascal VOC 2007* and *Caltech 101*. We used precomputed dense SIFT features for Pascal VOC 2007 and PHOW features [1] for Caltech 101.

On Caltech 101 dataset, we constructed three 5-class



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1: Active learning results on 13 Natural Scene Dataset. (a)-(b) are comparison results on two randomly sampled 5-class subsets. (c)-(e) are results on three randomly sampled 10-class subsets. (f) shows the comparison results of the proposed adaptive approach with the ones using different fixed $\beta$ values on the 5-class subset of (a).

subsets and two 10-class subsets, by randomly sampling from all the classes of the dataset. Figure 2a – Figure 2c show the average results over 10 runs on the three 5-class subsets for the five comparison methods, and Figure 2d – Figure 2e present the comparison results on the two 10-class subsets. On these subsets, different approaches have advantages in different scenarios, but the proposed adaptive active learning approach consistently outperforms all the others in all cases. Finally, we compared the proposed adaptive learning method and the non-adaptive versions of it with different fixed $\beta$ values on the 5-class subset of Figure 2c. The comparison results are reported in Figure 2f. Again, these results show the adaptive $\beta$ selection method outperforms the methods using fix $\beta$ values even within the same measure combination framework.

On the *Pascal VOC 2007 dataset*, we constructed two randomly selected 5-class subsets and one 10-class subset to conduct experiments. Figure 3 presents the average results produced on Pascal VOC 2007 dataset. The proposed approach again outperforms the other methods in all cases. The adaptive $\beta$ selection procedure produces obviously better results than using fixed $\beta$ values. Moreover, we also collected the class distribution information of the queried instances in one of the 5-class subsets, and depict it as a histogram in Figure 3e, which suggests the images from different classes are not equally informative.

In summary, the proposed adaptive active learning method demonstrates consistently superior performance to the other methods in both the scene recognition experiments and the object recognition experiments.

## 5. Conclusion

In this paper, we presented a new adaptive active learning approach which combines an information density measure with a most uncertainty measure together in an adaptive way to conduct instance selection. The adaptive combination procedure allows the proposed method to best integrate the strengths of the two measures in different stages and scenarios of active learning. This new approach can effectively use the information contained in the unlabeled data to improve the performance of uncertainty sampling. In our experiments on image classification problems, we showed that the proposed approach is able to shrink the training set required for learning a good classifier considerably, comparing to a number of existing active learning methods.

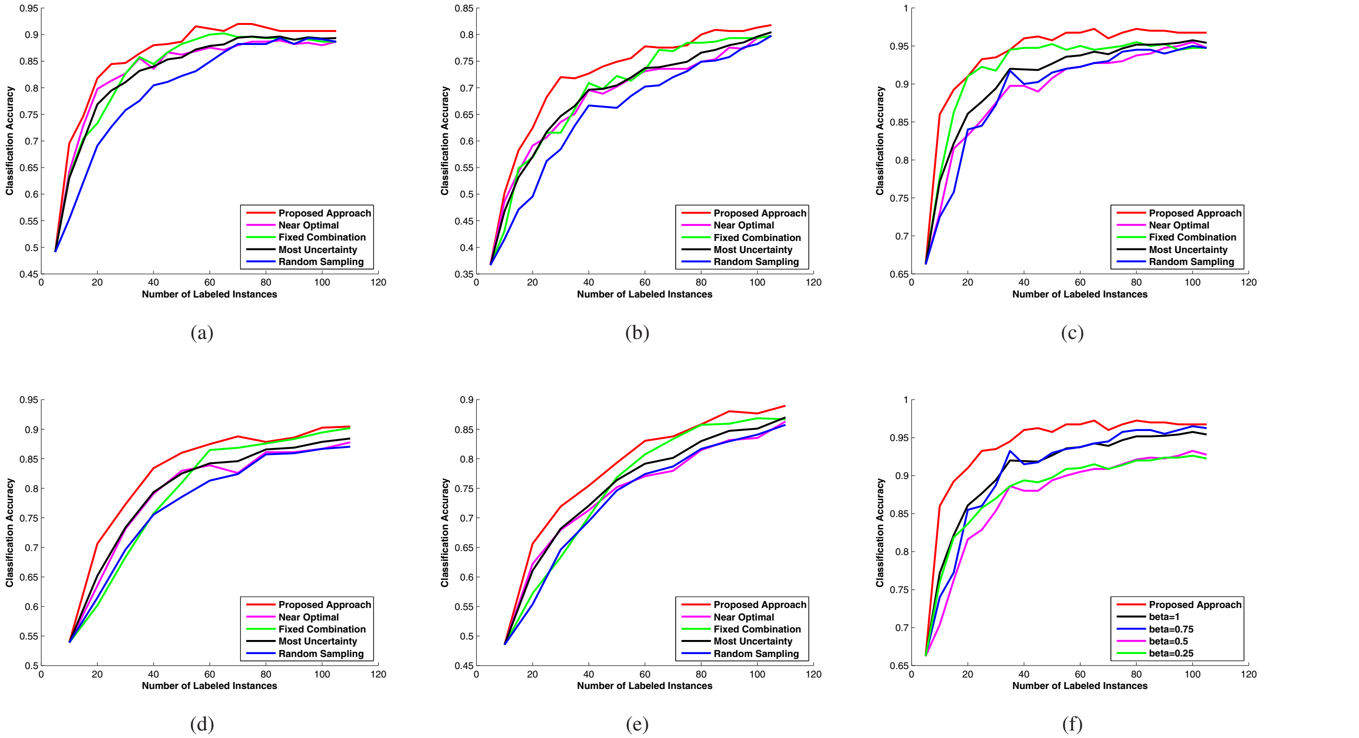


Figure 2: Active learning results on Caltech 101 Dataset. (a), (b), (c) are results on three randomly selected 5-class subsets. (d) and (e) are results on two randomly selected 10-class subsets. (f) gives the comparison results of adaptive active learning v.s. active learning with fixed $\beta$ values on the subset (c).

## References

[1] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[2] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *ICML*, 2000.

[3] S. Chakraborty, V. Balasubramanian, and S. Panchanathan. Dynamic batch mode active learning. In *CVPR*, 2011.

[4] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *JAIR*, 4, 1996.

[5] A. Corduneanu and T. Jaakkola. On information regularization. In *UAI*, 2003.

[6] P. Donmez, J. Carbonell, and P. Bennett. Dual strategy active learning. In *ECML*, 2007.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004.

[8] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[9] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28, 1997.

[10] C. Guestrin, A. Krause, and A. Singh. Near-optimal sensor placements in Gaussian processes. In *ICML*, 2005.

[11] Y. Guo and R. Greiner. Optimistic active learning using mutual information. In *IJCAI*, 2007.

[12] E. Hauptmann, W. Lin, R. Yan, J. Yang, and M. Chen. Extreme video retrieval: joint maximization of human and computer performance. In *Multimedia*, 2006.

[13] P. Jain and A. Kapoor. Active learning for large multi-class problems. In *CVPR*, 2009.

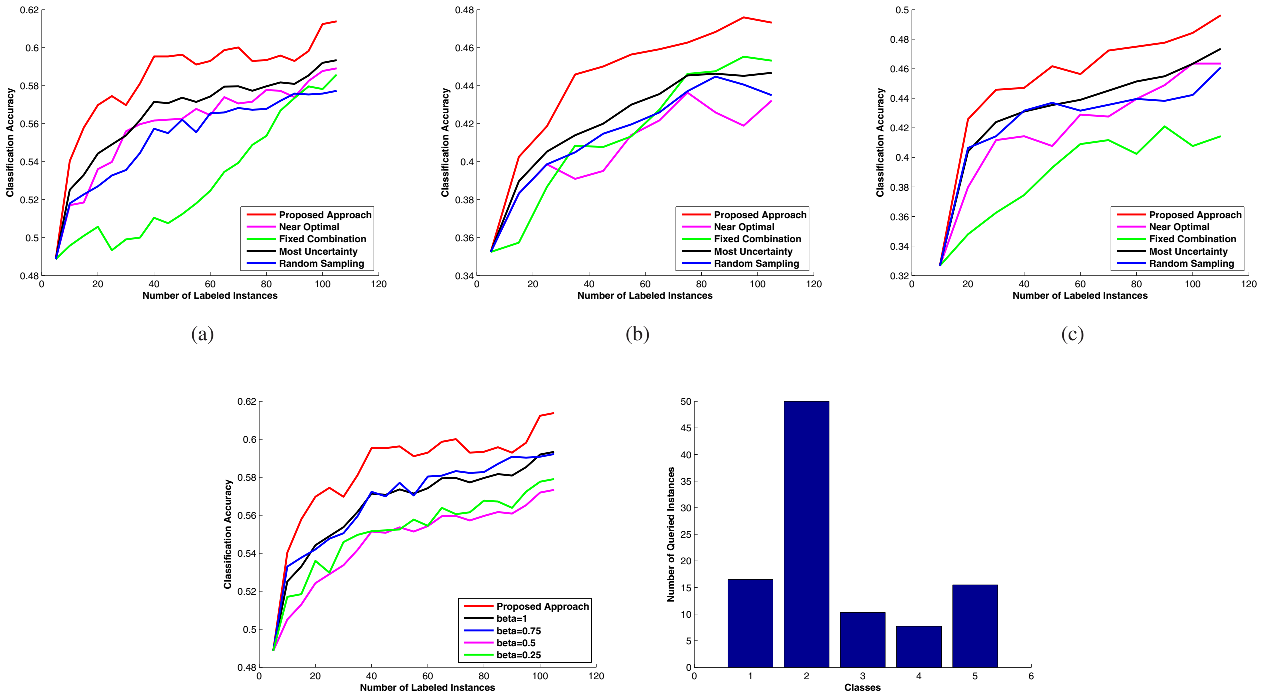
[14] A. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[15] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007.

[16] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker. Which faces to tag: Adding prior constraints into active learning. In *ICCV*, 2009.

[17] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *IPSN*, 2006.

[18] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Inter. ACM-SIGIR Conf. on Research and Develop. in Info. Retrieval*, 1994.



(a) (b) (c)

(d) (e)

Figure 3: Active learning results on VOC 2007. (a)-(b) are the results on two randomly chosen 5-class subsets. (c) shows the results on one randomly chosen 10-class subset. (d) demonstrates that adaptive $\beta$ is better than any fixed $\beta$. (e) shows the class distribution of queried instances in one of the 5-class problems.

[19] A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *ICML*, 1998.

[20] H. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML*, 2004.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.

[22] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang. Two-dimensional active learning for image classification. In *CVPR*, 2008.

[23] R. Rosales, P. Krishnamurthy, and B. Rao. Semi-supervised active learning for modeling medical concepts from free text. In *ICMLA*, 2007.

[24] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[25] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML*, 2000.

[26] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*, 2008.

[27] M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *NIPS*, 2002.

[28] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *ICML*, 2000.

[29] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011.

[30] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 2010.

[31] R. Yan, J. Yang, and A. Hauptmann. Automatically labeling video data using multi-class active learning. In *ICCV*, 2003.

[32] C. Zhang and T. Chen. An active learning framework for content-based information retrieval. *IEEE Trans on Multimedia*, 4:260–258, 2002.

[33] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.

[34] L. Zhang, Y. Tong, and Q. Ji. Active image labeling and its application to facial action labeling. In *ECCV*, 2008.

[35] T. Zhang and F. Oles. A probability analysis on the value of unlabeled data for classification problems. In *ICML*, 2000.

[36] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML Workshop on Continuum from Labeled to Unlabeled Data in Mach. Learn. and Data Mining*, 2003.