Two-Dimensional Active Learning for Image Classification

[†]Guo-Jun Qi, [‡]Xian-Sheng Hua, [‡]Yong Rui, [†]Jinhui Tang, [‡]Hong-Jiang Zhang
[†]MOE-Microsoft Key Laboratory of Multimedia Computing and Communication
& Department of Automation, University of Science and Technology of China
{qgj, jhtang}@mail.ustc.edu.cn
[‡]Microsoft Research Asia
{xshua, yongrui, hjzhang}@microsoft.com

Abstract

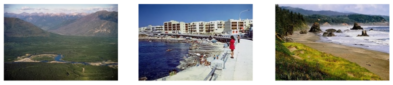
In this paper, we propose a two-dimensional active learning scheme and show its application in image classification. Traditional active learning methods select samples only along the sample dimension. While this is the right strategy in binary classification, it is sub-optimal for multi-label classification. In multi-label classification, we argue that, for each selected sample, only a part of more effective labels are necessary to be annotated while others can be inferred by exploring the correlations among the labels. The reason is that the contributions of different labels to minimizing the classification error are different due to the inherent label correlations. To this end, we propose to select sample-label pairs, rather than only samples, to minimize a multi-label Bayesian classification error bound. This new active learning strategy not only considers the sample dimension but also the label dimension, and we call it Two-Dimensional Active Learning (2DAL). We also show that the traditional active learning formulation is a special case of 2DAL when there is only one label. Extensive experiments conducted on two real-world applications show that the 2DAL significantly outperforms the best existing approaches which did not take label correlation into account.

1. Introduction

Image semantic understanding is typically formulated as either a multi-class or a multi-label classification problem [15][2]. In the multi-class setting, each image is classified into *one and only one* predefined category. In other words, only one label is assigned to an image in this setting. Real-world applications [2], however, require that one or multiple labels can be assigned to an image. This requirement results in multi-label classification, which is significantly more challenging, and will be the focus of this paper. Specifically, we will use active learning as the tool, and extend it from a *one-dimensional* sample-centric approach to a *two-dimensional* joint sample-label-centric approach for multi-label image classification.

Active learning is one of the most used methods in image classification, as it can significantly reduce the human cost in labeling training samples. Specifically, active learning methods iteratively annotate a set of elaborately selected samples so that the classification error is minimized in each iteration. As a result, the total number of training samples that need to be labeled is smaller than non active learning approaches. It is clear that the core of any active learning approach is the sample selection strategy. In the past decade, a number of active learning approaches were developed by using different sample selection strategies [14][4][10][8]. Most of these approaches focus on the binary or multi-class classification scenario [10][4][15]. However, in many real-world applications such as image and video retrieval [2][12], text search [16] and bioinformatics [6], a sample is usually associated with multiple labels rather than a single one. Under such a multi-label setting, each sample will be annotated as "positive" or "negative" for each and every label (See figure 1 for some examples). As a result, active learning with multi-labeled samples is much more challenging than that with binary-labeled ones, especially when the number of labels is large.

A direct way to tackle active learning under multi-label setting is to translate it into a set of binary problems, i.e., each category/label is independently handled by a binary active learning algorithm. For example, in [11][3] two research groups have proposed such a binary-based active learning algorithm for multi-labeled classification problem, respectively. However, this type of approaches does not take into account the inherent relationships among multiple labels. In this paper, we propose a novel active learning strategy which iteratively selects *sample-label pairs* to minimize



Field	P	N	N
Mountain	P	N	P
Urban	N	P	N
Beach	N	P	P
People	N	P	N

Figure 1. Some examples of multi-labeled images. "P" means positive label and "N" means negative label.

the expected classification error. Specifically, in each iteration, the annotators are only required to annotate/confirm a selected part of labels of selected samples while the remaining unlabeled part will be inferred according to the label correlations. We call this strategy *2 Dimensional Active Learning* (2DAL) since it considers not only the samples to be labeled along the *sample dimension* but also the labels associated with these samples along the *label dimension*.

An intuitive explanation of this strategy is that there exist both *sample* and *label redundancies* for multi-labeled samples. Therefore, annotating a set of selected sample-label pairs provides enough information for training the classifiers since the information in the selected pairs can be propagated to the rest along both the *sample* and the *label* "dimensions". Unlike existing binary-based active learning strategies [11][3] which only take the sample redundancy into account, the 2DAL strategy additionally considers the *label* dimension to leverage the rich relationships embedded in the multiple labels. 2DAL efficiently selects an informative part of the labels rather than all the labels for a particular selected sample. Such a strategy significantly reduces the required human labors during active learning. For example, *Field* and *Mountain* tend to occur simultaneously in an image. Therefore, it is reasonable to select only one label (e.g., *Mountain*) for annotation since the uncertainty of the other label can be remarkably decreased after annotating this one. Another example is *Mountain* and *Urban*, in contrast to *Field* and *Mountain*, these two labels often do not occur simultaneously. Thus, annotating one of them most likely will probably eliminate the existence of the other.

To realize 2DAL, we will answer the following questions in this paper:

1 What is the proper selection strategy for finding the sample-label pairs? To address this issue, we formulate the selection of sample-label pairs as minimizing a derived *Multi-Label Bayesian Classification Error Bound*. We will demonstrate that selecting sample-label pairs in this way will significantly reduce the uncertainty of both the samples and the labels.

2 How can we model the label relationships/*correlations*? Since the proposed 2DAL strategy utilizes the label dependencies to reduce labeling cost, the underlying classifier should be able to model the corresponding label correlations. Accordingly, we propose a *Kernelized Maximum Entropy Model* (KMEM) to model such correlations. Furthermore, since the 2DAL strategy only annotates a sub set of labels, we formulate an *Expectation-Maximum*(EM) [5] algorithm to solve the incomplete labeling problem.

To the best of our knowledge, we are the first to present the study of active learning on the granularity of sample-label pairs, with both theoretical analysis and empirical results on real-world data sets. The rest of the paper is organized as follows. In section 2, we present the 2DAL selection strategy used in the proposed active learning algorithm. We also show that the traditional active learning formulation is a special case of 2DAL when there is only one lable. After that, a Kernelized Maximum Entropy Model is proposed in section 3 to model the label correlations. In addition, an Expectation-Maximum (EM) algorithm is also given in this section to solve the incomplete labeling problem. In section 4 we evaluate the proposed 2DAL with comparison with the state-of-the-art one dimensional active learning approach on two real-world data sets. Finally, we conclude in section 5.

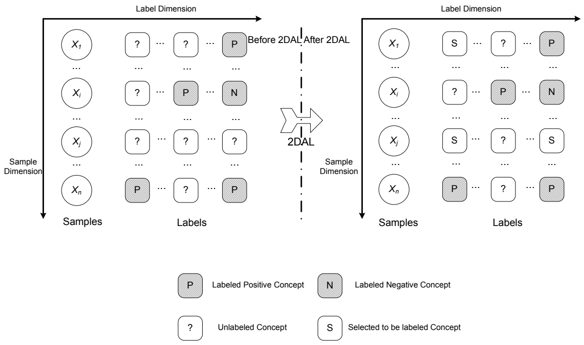


Figure 2. Proposed two-dimension 2DAL strategy

2. Two-Dimensional Active Learning

In this section, we start by detailing the underlying idea of the proposed 2DAL strategy in multi-label setting from the perspective of information theory. Then, a Bayesian error bound is derived that states the expected classification error given a selected sample-label pair. The proposed 2DAL strategy will then be deduced by selecting the sample-label pairs which optimize this bound.

2.1. The proposed 2DAL strategy

Figure 2 illustrates the proposed 2DAL strategy. Different from the typical binary active learning formulation that selects the most informative samples for annotation, we jointly select both the *samples* and *labels* simultaneously. The underlying assumption is that different labels of a certain sample have different contributions to minimizing the expected classification error of the to-be-trained classifier. And annotating a portion of well-selected labels may provide sufficient information for learning the classifier. As shown in Figure 1, this strategy trades off between the annotation labors and the learning performance along two dimensions, i.e., the *sample* and *label* dimensions. In essence, the multi-label classifiers do have *uncertainty* along different labels as well as different samples. Traditional active learning algorithms can be seen as a *one-dimension* active selection approach, which only reduces the *sample uncertainty*. In contrast, 2DAL is a *two-dimensional* active learning strategy, which selects the most "informative" sample-label pairs to reduce the uncertainty along the dimensionalities of both *sample* and *label*. More specifically, along *label* dimension all of the labels correlatively interact. Therefore, once partial labels are annotated, the left unlabeled concepts can then be inferred based on label correlations. Theoretically, the label correlations have a connection with the expected Bayesian Error Bound (see the following lemma and theorem in section 2.2), and thus these label correlations can help to reduce the prediction errors in the testing set during the active learning procedure. This approach saves much labor from fully annotating multiple labels. Thus, it is far more efficient when the number of labels is huge. For instance, an image may be associated with thousands and hundreds of concepts. That means a full annotation strategy will pay large labor costs for only one image. On the contrary, 2DAL only selects the most informative labels for annotation. In the following section, we will derive such a two-dimension selection criterion based on a derived *Bayesian classification error bound* in multi-label setting.

On the other hand, it is worth noting that as illustrated in Figure 2, during the learning process, some samples may be lack of some labels since only a partial of labels are annotated. This is different from traditional active learning algorithm. In the section 3.2, we will address how to learn the classification model from incomplete labels.

2.2. Multi-labeled Bayesian error bound

The 2DAL learner requests annotations on the basis of sample-label pairs which, once incorporated into the training set, are expected to result in the lowest generalization error. Here we will first derive a *Multi-Labeled Bayesian Error Bound* when a selected sample-label pair is labeled under multi-label setting, and 2DAL accordingly will iteratively select the ones to minimize this bound.

Before we move further, we first define some notations. For each sample $\boldsymbol{x}$, it has m labels $y_i (1 \leq i \leq m)$ and each of them indicates whether its corresponding semantic concept occurs. As stated before, in each 2DAL iteration, some of these labels have already been annotated while others not. Let $U(\boldsymbol{x}) \triangleq \{i | (\boldsymbol{x}, y_i)$ is unlabeled sample-label pair.$\}$ denote the set of indices of unlabeled part and $L(\boldsymbol{x}) \triangleq \{i | (\boldsymbol{x}, y_i)$ is labeled sample-label pair.$\}$ denote the labeled part. Note that $L(\boldsymbol{x})$ can be an empty set $\varnothing$, which indicates that no label has been annotated for $\boldsymbol{x}$. Let $P(\boldsymbol{y}|\boldsymbol{x})$ be the conditional distribution over samples, where $\boldsymbol{y} = \{0, 1\}^m$ is the complete label vector and $P(\boldsymbol{x})$ be the marginal sample distribution.

First, we establish a Bayesian error bound for classifying one unlabeled y_i once y_s is actively selected for annotating. This error bound originates from the equivocation bound given in [7], and we extend it to multi-label setting so it can handle sample-label pairs.

Lemma 1. *Given a sample* $\boldsymbol{x}$ *and its labeled and unlabeled parts* $U(\boldsymbol{x})$ *and* $L(\boldsymbol{x})$*. Once* y_s *is selected to ask for labeling (but not yet annotated), the Bayesian classification error* $\mathcal{E}\left(y_i | y_s, y_{L(\boldsymbol{x})}, \boldsymbol{x}\right)$ *for an unlabeled* $y_i, i \in U(\boldsymbol{x})$ *is bounded as*

$$\begin{aligned} &\tfrac{1}{2} H\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) - \epsilon \leq \mathcal{E}\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\ &\leq \tfrac{1}{2} H\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \end{aligned} \tag{1}$$

where

$$\begin{aligned} &H\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) = \sum_{t,r \in \{0,1\}} \{-P\left(y_i = t, y_s = r | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\ &\times \log P\left(y_i = t | y_s = r; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right)\} \end{aligned}$$

is the conditional entropy of y_i *given the selected part* y_s *(both* y_i *and* y_s *are random variables since they have not been labeled) and* $y_{L(\boldsymbol{x})}$ *is the known labeled part;* $\epsilon = \frac{1}{2} \log \frac{5}{4}$ *is a constant.*

This lemma will be proven in the appendix.

Remark 1. *It is worth noting that this bound is irrelevant to the true label of the selected* y_s*. In fact, before the annotator gives the label of* y_s*, the true value of* y_s *is unknown. However, no matter what* y_s *holds,* 1 *or* 0*, this bound always holds.*

Based on this lemma, we can obtain the following theorem which bounds the multi-label error:

Theorem 1. *(Multi-labeled Bayesian classification error bound) Under the condition of lemma 1, the Bayesian classification error bound* $\mathcal{E}(\boldsymbol{y} | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x})$ *for sample* $\boldsymbol{x}$ *over all the labels* $\boldsymbol{y}$ *is*

$$\begin{aligned} &\mathcal{E}\left(\boldsymbol{y} | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\ &\leq \tfrac{1}{2m} \textstyle\sum_{i=1}^{m} \left\{H\left(y_i | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) - MI\left(y_i; y_s | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right)\right\} \end{aligned} \tag{2}$$

where $MI(y_i; y_s | y_{L(\boldsymbol{x})}, \boldsymbol{x})$ *is the mutual information between the random variables* y_i *and* y_s *given the known labeled part* $y_{L(\boldsymbol{x})}$*.*

Proof.

$$\begin{aligned} &\mathcal{E}\left(\boldsymbol{y} | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\ &\overset{(1)}{=} \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \mathcal{E}\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\ &\overset{(2)}{\leq} \tfrac{1}{2m} \textstyle\sum_{i=1}^{m} H\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\ &\overset{(3)}{=} \tfrac{1}{2m} \textstyle\sum_{i=1}^{m} \left\{H\left(y_i | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) - MI\left(y_i; y_s | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right)\right\} \end{aligned} \tag{3}$$

where (2) directly comes from Lemma 1, (3) makes use of the relationship between mutual information and entropy: $MI(X;Y) = H(X) - H(X|Y)$. □

We are concerned with *pool-based active learning*, i.e., a large pool $\mathcal{P}$ is available to the learner sampled from $P(\boldsymbol{x})$ and the proposed 2DAL then selects the most informative sample-label pairs from the pool. We first write the expected Bayesian classification error over all samples in $\mathcal{P}$ before selecting a sample-label pair $(\boldsymbol{x}_s, y_s)$

$$\mathcal{E}^b\left(\mathcal{P}\right) = \frac{1}{|\mathcal{P}|} \textstyle\sum_{\boldsymbol{x} \in \mathcal{P}} \mathcal{E}\left(\boldsymbol{y} | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \tag{4}$$

We can use the above classification error on the pool to estimate the expected error over the full distribution $P(\boldsymbol{x})$, i.e., $E_{P(\boldsymbol{x})}\mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x})},\boldsymbol{x}\right) = \int P(\boldsymbol{x})\mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)d\boldsymbol{x}$, because the pool not only provides a finite set of samples but also an estimation of $P(\boldsymbol{x})$. After selecting the pair $(\boldsymbol{x}_s, y_s)$, the expected Bayesian classification error over the pool $\mathcal{P}$ is

$$\begin{aligned}&\mathcal{E}^a\left(\mathcal{P}\right)\\&= \tfrac{1}{|\mathcal{P}|}\left\{\mathcal{E}\left(\boldsymbol{y}|y_s;y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right) + \textstyle\sum_{\boldsymbol{x}\in\mathcal{P}\backslash\boldsymbol{x}_s}\mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\right\}\\&= \tfrac{1}{|\mathcal{P}|}\{\mathcal{E}\left(\boldsymbol{y}|y_s;y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right) - \mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&+\textstyle\sum_{x\in\mathcal{P}}\mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\}\end{aligned} \tag{5}$$

Therefore, the reduction of the expected Bayesian classification after selecting $(\boldsymbol{x}_s, y_s)$ over the whole pool $\mathcal{P}$ is

$$\Delta\mathcal{E}\left(\mathcal{P}\right) = \mathcal{E}^b\left(\mathcal{P}\right) - \mathcal{E}^a\left(\mathcal{P}\right) \tag{6}$$

Thus our goal is to select a best $(\boldsymbol{x}_s^*, y_s^*)$ to maximize the above expected error reduction. That is,

$$\begin{aligned}&(\boldsymbol{x}_s^*, y_s^*) = \arg\max\nolimits_{\boldsymbol{x}_s\in\mathcal{P},y_s\in U(\boldsymbol{x}_s)}\Delta\mathcal{E}\left(\mathcal{P}\right)\\&= \arg\min\nolimits_{\boldsymbol{x}_s\in\mathcal{P},y_s\in U(\boldsymbol{x}_s)} -\Delta\mathcal{E}\left(\mathcal{P}\right)\end{aligned} \tag{7}$$

Applying Lemma 1 and Theorem 1, we have

$$\begin{aligned}&-\Delta\mathcal{E}\left(\mathcal{P}\right) = \mathcal{E}^a\left(\mathcal{P}\right) - \mathcal{E}^b\left(\mathcal{P}\right)\\&\overset{(1)}{=} \tfrac{1}{|\mathcal{P}|}\{\mathcal{E}\left(\boldsymbol{y}|y_s;y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right) - \mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&+\textstyle\sum_{\boldsymbol{x}\in\mathcal{P}}\mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\} - \tfrac{1}{|\mathcal{P}|}\textstyle\sum_{\boldsymbol{x}\in\mathcal{P}}\mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&= \tfrac{1}{|\mathcal{P}|}\left\{\mathcal{E}\left(\boldsymbol{y}|y_s;y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right) - \mathcal{E}\left(\boldsymbol{y}|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\right\}\\&\overset{(2)}{\leq} \tfrac{1}{|\mathcal{P}|}\{\tfrac{1}{2m}\textstyle\sum_{i=1}^m H\left(y_i|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&-\tfrac{1}{2m}\textstyle\sum_{i=1}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&-\tfrac{1}{m}\textstyle\sum_{i=1}^m\mathcal{E}\left(y_i|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\}\\&\overset{(3)}{\leq} \tfrac{1}{|\mathcal{P}|}\{\tfrac{1}{2m}\textstyle\sum_{i=1}^m H\left(y_i|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&-\tfrac{1}{2m}\textstyle\sum_{i=1}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&-\tfrac{1}{m}\textstyle\sum_{i=1}^m\left(\tfrac{1}{2}H\left(y_i|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)-\epsilon\right)\}\\&= \tfrac{1}{|\mathcal{P}|}\left\{\epsilon - \tfrac{1}{2m}\textstyle\sum_{i=1}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\right\}\end{aligned} \tag{8}$$

The equality (1) comes from Eqn. 4 5 . The first inequality (2) follows the Theorem 1 and the second inequality (3) comes from the lower bound of Lemma 1.

Consequently, by minimizing the obtained Bayesian error bound 8, we can select the sample-label pair for annotation according to

$$\begin{aligned}&(\boldsymbol{x}_s^*, y_s^*)\\&= \mathop{\arg\min}_{\boldsymbol{x}_s\in\mathcal{P},y_s\in U(\boldsymbol{x}_s)}\tfrac{1}{|\mathcal{P}|}\left\{\epsilon - \tfrac{1}{2m}\sum_{i=1}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\right\}\\&= \mathop{\arg\max}_{\boldsymbol{x}_s\in\mathcal{P},y_s\in U(\boldsymbol{x}_s)}\textstyle\sum_{i=1}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\end{aligned} \tag{9}$$

2.3. Further Discussions

1 As we discussed in section 2.1, the proposed 2DAL approach is an active learning algorithm along two dimensions, which reduces not only *sample uncertainty* but also *label uncertainty*. The above selection strategy Eqn. 9 actually well-reflects these two targets. The last term in Eqn. 9 can be rewritten as

$$\begin{aligned}&\textstyle\sum_{i=1}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&= MI\left(y_s;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)+\sum_{i=1,i\neq s}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\\&= H\left(y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right) + \sum_{i=1,i\neq s}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)\end{aligned} \tag{10}$$

As we can see, the objective selection function for 2DAL has been divided into two parts: $H\left(y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)$ and $\sum\limits_{i=1,i\neq s}^m MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)$. The former entropy measures the uncertainty of the selected pair $(\boldsymbol{x}_s^*, y_s^*)$ itself, and this is consistent with the typical one dimensional active learning algorithm, i.e., to select the most uncertain samples near the classification boundary [10][9]. On the other hand, the latter mutual information term measures the statistical redundancy among the selected label and the rest. By maximizing these mutual information terms, 2DAL can provide information for the inference of other labels to help reduce their label uncertainty. Therefore, the obtained strategy confirms our motivation of selecting the most informative sample-label pairs to reduce the uncertainties along both *sample* and *label* dimension. Note that when there is only one label for each sample, Eqn. 10 reduces to $H(y_s|\boldsymbol{x}_s)$. The selection criterion becomes the same as the traditional binary-based criterion, i.e., to select the most uncertain sample for annotation [9] [14].

2 When computing the mutual information terms in Eqn. 9, we need the distribution $P(\boldsymbol{y}|\boldsymbol{x})$. However, the true distribution is unknown, but we can estimate it using the current learner. As stated in [13], such an approximation is reasonable because the most useful labeling is usually consistent with the learner's prior belief over the majority (but not all) of the unlabeled pairs.

3 It is worth indicating that the posterior $P(\boldsymbol{y}|\boldsymbol{x})$ is significant in modeling the label correlations. If we assume the independence among the different labels, i.e., $P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^m P(y_i|\boldsymbol{x})$ and correspondingly the mutual information term will become $MI\left(y_i;y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right) = 0, i \neq s$. In this case, the selection criterion reduces to $(\boldsymbol{x}_s^*, y_s^*) = \arg\max_{\boldsymbol{x}_s\in\mathcal{P},y_s\in U(\boldsymbol{x}_s)} H\left(y_s|y_{L(\boldsymbol{x}_s)},\boldsymbol{x}_s\right)$, that is, to select the most uncertain sample-label pair. Such a criterion neglects the label correlations and will be less efficient in reducing label uncertainty. Therefore, a statistical method that can model the label correlations is required to adopt. We introduce such a Bayesian model in the following section.

3. Maximum Entropy Model and EM Variant

In the above 2DAL strategy, we have indicated that a statistical model is needed to measure label correlations. However, common multi-label classifiers, such as one-against-rest encoded binary SVM and others, tackle the classification of multi-labeled samples in an independent manner. These models neglect the label correlations and, hence, do not fit our target. In this section, we will introduce a multi-labeled Bayesian classifier in which the relations among different labels are well modeled.

3.1. Kernelized maximum entropy model

The principle of *Maximum Entropy Model* (MEM) is to model all known, and assume nothing about the unknown. Traditional single-label data classification suffers from the same problem as binary SVM. [16] extends the single labeled MEM to multi-labeled case. This model is linear and can be effective on a set of samples that vary linearly. However, it will fail to capture the structure of the feature space if the variations among the samples are nonlinear. But image classification is actually in this case when one is trying to extract features from image categories that vary in their appearance, illumination conditions and complex background clutters. Therefore, a nonlinear version of such a MEM is required to classify the images based on their nonlinear feature structure. Moreover, they do not address the problem brought by incomplete labels. We first introduce the model in [16] and further extend it to a nonlinear case by incorporating a kernel function into the model. Such an extension is used as the underlying classifier in 2DAL.

Let $\widetilde{Q}(\boldsymbol{x},\boldsymbol{y}), Q(\boldsymbol{x},\boldsymbol{y})$ denote the empirical and the model distribution, respectively. The optimal multi-label model can be obtained by solving the following formulation [16]:

$$\begin{array}{l}
\hat{P} = \arg\max_P H(\boldsymbol{x},\boldsymbol{y}|Q) = \arg\min_P \langle \log P(\boldsymbol{y}|\boldsymbol{x}) \rangle_Q \\
s.t. \langle y_i \rangle_Q = \langle y_i \rangle_{\tilde{Q}} + \eta_i, \\
\langle y_i y_j \rangle_Q = \langle y_i y_j \rangle_{\tilde{Q}} + \theta_{il}, 1 \leq i < j \leq m \\
\langle y_i x_l \rangle_Q = \langle y_i x_l \rangle_{\tilde{Q}} + \phi_{il}, 1 \leq i \leq m, 1 \leq l \leq d; \\
\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) = 1
\end{array} \tag{11}$$

where $H(\boldsymbol{x},\boldsymbol{y}|Q)$ is the entropy of $\boldsymbol{x}$ and $\boldsymbol{y}$ given distribution Q, and $\langle\cdot\rangle_Q$ denotes the expectation with respect to distribution Q. d is the dimension of the feature vector $\boldsymbol{x}$ and x_l represents its l-th element. η_i, θ_{il} and ϕ_{il} are the estimation errors following the Gaussian distribution which serve to smooth the MEM to improve the model's generalization ability. By modeling the pair-wise label correlations, the obtained model reveals the underlying label correlations. Formulation 11 can be solved by Lagrange Multiplier algorithms and the obtained posterior probability is $\hat{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{y}^T(\boldsymbol{b} + R\boldsymbol{y} + W\boldsymbol{x})\right)$, where $Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \boldsymbol{y}^T(\boldsymbol{b} + R\boldsymbol{y} + W\boldsymbol{x})$ is the partition function, and the parameters $\boldsymbol{b}$, W, and R are Lagrangian multipliers that need to be determined. The optimal parameters can be found by minimizing the Lagrangian:

$$\begin{array}{l}
\mathcal{L}(\boldsymbol{b}, R, W) = \left\langle -\log \hat{P}(\boldsymbol{y}|\boldsymbol{x}) \right\rangle_{\tilde{Q}} \\
+\frac{\lambda_b}{2n}||\boldsymbol{b}||_2^2 + \frac{\lambda_R}{2n}||R||_F^2 + \frac{\lambda_W}{2n}||W||_F^2 \\
= \left\langle -\boldsymbol{y}^T(\boldsymbol{b} + R\boldsymbol{y} + W\boldsymbol{x}) + \log Z(\boldsymbol{x}) \right\rangle_{\tilde{Q}} \\
+\frac{\lambda_b}{2n}||\boldsymbol{b}||_2^2 + \frac{\lambda_R}{2n}||R||_F^2 + \frac{\lambda_W}{2n}||W||_F^2
\end{array} \tag{12}$$

where $||.||_F$ denotes Frobenius norm and n is the number of samples in training set.

Now, we extend the above multi-labeled MEM to a nonlinear one so that the powerful kernel method can be adopted. A transformation ψ maps samples into a target space in which kernel function $k(\boldsymbol{x}', \boldsymbol{x})$ gives the inner product. We can rewrite the multi-labeled MEM as $\hat{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{y}^T(\boldsymbol{b} + R\boldsymbol{y}) + \boldsymbol{y}^T K(W, \boldsymbol{x})\right)$. According to the Representer Theorem, the optimal weighting vector of the single-labeled problem is a linear combination of samples. In the proposed multi-labeled setting, the mapped weighting matrix $\psi(W)$ can still be written as a linear combination of $\psi(\boldsymbol{x}_i)$ except that the combination coefficients are vectors instead of scalars, i.e.

$$\begin{array}{l}
\psi(W) = \sum_{i=1}^n \theta(\boldsymbol{x}_i)\psi^T(\boldsymbol{x}_i) \\
= \left[\begin{array}{cccc} \theta(\boldsymbol{x}_1) & \theta(\boldsymbol{x}_2) & \cdots & \theta(\boldsymbol{x}_n) \end{array}\right] \\
\cdot \left[\begin{array}{cccc} \psi(\boldsymbol{x}_1) & \psi(\boldsymbol{x}_2) & \cdots & \psi(\boldsymbol{x}_n) \end{array}\right]^T \\
= \Theta \cdot \left[\begin{array}{cccc} \psi(\boldsymbol{x}_1) & \psi(\boldsymbol{x}_2) & \cdots & \psi(\boldsymbol{x}_n) \end{array}\right]^T
\end{array} \tag{13}$$

where the summation is taken over the samples in the training set $\{\boldsymbol{x}_i\}_{i=1}^n$. $\theta(\boldsymbol{x}_i)$ is a $m \times 1$ coefficient vector and Θ is a $m \times n$ matrix in which each row is the weighting coefficients for each label. Accordingly, we have

$$\begin{array}{l}
K(W, \boldsymbol{x}) = \psi(W) \cdot \psi(\boldsymbol{x}) \\
= \Theta \cdot \left[\begin{array}{ccc} k(\boldsymbol{x}_1, \boldsymbol{x}) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}) \end{array}\right]^T = \Theta \cdot k(\boldsymbol{x})
\end{array} \tag{14}$$

and so

$$\begin{array}{l}
\hat{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{y}^T(b + \boldsymbol{R}\boldsymbol{y}) + \boldsymbol{y}^T k(W, \boldsymbol{x})\right) \\
= \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{y}^T(b + R\boldsymbol{y} + \Theta k(\boldsymbol{x}))\right)
\end{array} \tag{15}$$

where $k(\boldsymbol{x}) = \left[\begin{array}{ccc} k(\boldsymbol{x}_1, \boldsymbol{x}) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}) \end{array}\right]^T$ is a $n \times 1$ vector and it can be seen as a new representation of sample $\boldsymbol{x}$. Correspondingly, with the identity $||\psi(W)||_F^2 = tr(\psi(W)\psi(W)^T) = tr(\Theta K \Theta^T)$ the Lagrangian function Eqn. 12 can be rewritten as

$$\begin{array}{l}
\mathcal{L}(b, R, \Theta) = \left\langle -\log \hat{P}(\boldsymbol{y}|\boldsymbol{x}) \right\rangle_{\tilde{Q}} \\
+\frac{\lambda_b}{2n}||b||_2^2 + \frac{\lambda_R}{2n}||R||_F^2 + \frac{\lambda_W}{2n} tr(\Theta K \Theta^T)
\end{array} \tag{16}$$

where $K = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{n \times n}$ is the kernel matrix. We call the above model *Kernelized Maximum Entropy Model* (KMEM) in this paper. By minimizing Eqn.16, we can estimate the optimal parameters in KMEM.

3.2. EM based approach for incomplete labels

Given the partially labeled training set constructed by 2DAL (see Figure 2), we can handle the incomplete labels by integrating out the unlabeled part to yield the marginal distribution of the labeled part $\hat{P}(\boldsymbol{y}_{L(\boldsymbol{x})}|\boldsymbol{x}) = \sum_{\boldsymbol{y}_{U(\boldsymbol{x})}} \hat{P}(\boldsymbol{y}_{U(\boldsymbol{x})}, \boldsymbol{y}_{L(\boldsymbol{x})}|\boldsymbol{x})$. Then substitute it for $\hat{P}(\boldsymbol{y}|\boldsymbol{x})$ in Eqn. 16, we can obtain:

$$\begin{aligned}&\mathcal{L}(b, R, \Theta) = \left\langle -\log \textstyle\sum_{y_{U(\boldsymbol{x})}} \hat{P}(y_{U(\boldsymbol{x})}, y_{L(\boldsymbol{x})}|\boldsymbol{x}) \right\rangle_{\tilde{Q}} \\ &+ \tfrac{\lambda_b}{2n}||b||_2^2 + \tfrac{\lambda_R}{2n}||R||_F^2 + \tfrac{\lambda_W}{2n} tr(\Theta K \Theta^T)\end{aligned} \tag{17}$$

By minimizing Eqn. 17, we can find the optimal parameters for KMEM. However, it is difficult to minimize it directly. Instead, we use the *Expectation Maximization* (EM) algorithm [5] to solve this optimization problem:

E-Step: Given the current t-th step parameter estimation $\boldsymbol{b}_t, R_t, \Theta_t$, the $\mathcal{T}$-function (i.e., the expectation of the Lagrangain Eq. 16 under the current parameters given the labeled part) can be written as

$$\begin{aligned}&\mathcal{T}(b, R, \Theta|b_t, R_t, \Theta_t) \\ &= \left\langle -E_{U(\boldsymbol{x})|L(\boldsymbol{x});b_t,R_t,\Theta_t} \log \hat{P}(y_{U(\boldsymbol{x})}, y_{L(\boldsymbol{x})}|\boldsymbol{x}; b, R, \Theta) \right\rangle_{\tilde{Q}} \\ &+ \tfrac{\lambda_b}{2n}||b||_2^2 + \tfrac{\lambda_R}{2n}||R||_F^2 + \tfrac{\lambda_W}{2n} tr(\Theta K \Theta^T)\end{aligned} \tag{18}$$

where $E_{U(\boldsymbol{x})|L(\boldsymbol{x});\boldsymbol{b}_t,R_t,\Theta_t}$ is the expectation operator given the current estimated conditional probability $\hat{P}(\boldsymbol{y}_{U(\boldsymbol{x})}|\boldsymbol{y}_{L(\boldsymbol{x})}, \boldsymbol{x}; \boldsymbol{b}_t, R_t, \Theta_t)$.

M-Step: Update the parameters by minimizing $\mathcal{T}$-function:

$$\boldsymbol{b}_{t+1}, R_{t+1}, \Theta_{t+1} = \arg \min_{\boldsymbol{b},R,\Theta} \mathcal{T}(\boldsymbol{b}, R, \Theta|\boldsymbol{b}_t, R_t, \Theta_t) \tag{19}$$

The derivatives of $\mathcal{T}$-function with respect to its parameters $\boldsymbol{b}, R, \Theta$ is

$$\begin{aligned}&\tfrac{\partial \mathcal{T}}{\partial b_i} = \langle y_i \rangle_Q - \left\langle E_{y_i|L(\boldsymbol{x});b_t,R_t,\Theta_t} y_i \right\rangle_{\tilde{Q}} + \tfrac{\lambda_b}{n} b_i \\ &\tfrac{\partial \mathcal{T}}{\partial R_{ij}} = \langle y_i y_j \rangle_Q - \left\langle E_{y_i,y_j|L(\boldsymbol{x});b_t,R_t,\Theta_t} y_i y_j \right\rangle_{\tilde{Q}} + \tfrac{\lambda_R}{n} R_{ij} \\ &\tfrac{\partial \mathcal{T}}{\partial \Theta_{il}} = \langle y_i k(\boldsymbol{x}_l, \boldsymbol{x}) \rangle_Q - \left\langle E_{y_i|L(\boldsymbol{x});b_t,R_t,\Theta_t} y_i k(\boldsymbol{x}_l, \boldsymbol{x}) \right\rangle_{\tilde{Q}} \\ &+ \tfrac{\lambda_W}{n} \textstyle\sum_{k=1}^{n} \Theta_{ik} k(\boldsymbol{x}_k, \boldsymbol{x}_l)\end{aligned} \tag{20}$$

Given these derivatives, we can use the efficient gradient descent methods (e.g., LMVM [1]) to minimize Eqn. 18.

4. Experiments

In this section, we will evaluate the proposed 2DAL strategy on two real-world used data sets. The first data set is a natural scene set with six image categories. The second is a biological data set with 14 different types of genes. These two data sets are publicly available at http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/. We compare the proposed 2DAL with the state-of-the-art active learning approaches.

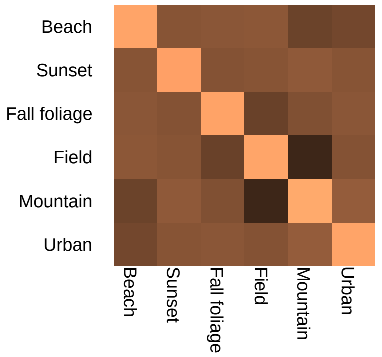


Figure 3. The mutual information between different concepts in Scene data set

Class	Total	Class	Total
Beach	369	Beach+Mountain	38
Sunset	364	Foliage+Mountain	13
Foliage	360	Field+Mountain	75
Field	327	Field+Foliage+Mountain	1
Beach+Field	1	Urban	405
Foliage+Field	23	Beach+Urban	19
Mountain	405		

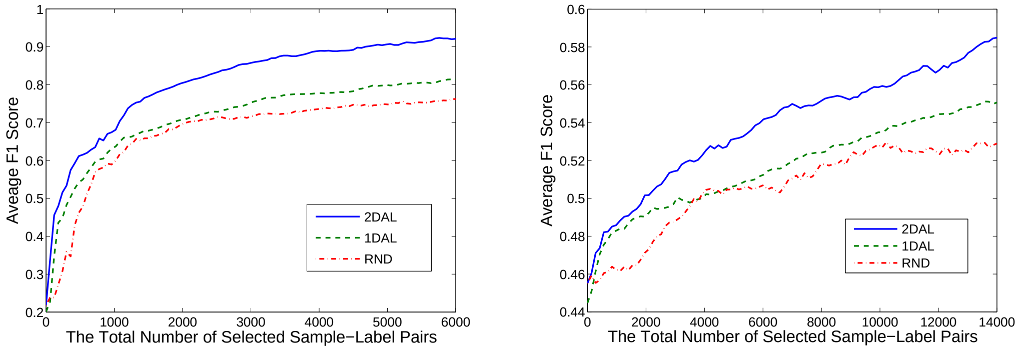
Table 1. The description about the *Scene* data set

4.1. Natural scene data set

This natural scene data set is first used in a previous research on the multi-labeled image scene classification problem [2]. It contains 2,407 natural images belonging to one or more of six natural scene categories including beach, sunset, fall foliage, field, mountain, and urban. Since the data sets are multi-labeled, there are 14,442 sample-label pairs in this set. Each sample in this set has been assigned by three positive labels at most. Table 1 describes the multi-label distribution in this set. We can see that 177 samples have more than one positive labels. Although this number is not large, it does not say the label correlation is low. In fact, the statistical correlations between the labels are determined by not only the correlations between positive labels but also those between the negative labels, as well as between positive and negative ones. In figure 3, we illustrate the mutual information calculated over the whole data set. According to the information theory, the mutual information considers all kinds of the correlations among the positive/negative labels as stated above. From this illustration, the correlations between the labels are obvious. Note that, the mutual information computed here is not the one used in 2DAL as Eqn. 9. In Eqn. 9, the mutual information is calculated from the statistical model KMEM.

For the features used in this experiment, an image is first converted into CIE Luv color space and then the first and second color moments (mean and variance) are extracted over a 7×7 grid on the image. The end result is a $49 \times 2 \times 3 = 294$ dimension feature vector [2].

In this experiment, we compare the following three active learning strategies:



(a) *Scene*

(b) *Yeast*

Figure 4. The performance of five active learning strategies over two real-world data sets (a)*Scene* (b)*Yeast*

Class	2DAL	1DAL	RND
Beach	0.9523	0.8652	0.6744
Sunset	0.9916	0.9421	0.9002
Fall Foliage	0.9887	0.9338	0.8927
Field	0.9588	0.8813	0.8071
Mountain	0.7806	0.6457	0.6122
Urban	0.8534	0.6162	0.6856

Table 2. $F1$ scores after 100 iterations on six scene categories.

1 The proposed 2DAL strategy: using the proposed sample-label pair selection criterion in Section 2.2, with KMEM as the underlying classifier.

2 1D active learning strategy (1DAL): using the mean-max loss active learning strategy that has been proposed in the previous work [11] on multi-label active learning. As stated in Section 1, this strategy selects only along the sample dimension. It does not take advantages of the label correlations to reduce human labeling cost. Therefore when one sample is selected, all its labels have to be labeled.

3 Random strategy (RND): selecting the sample-label pairs at random. For the sake of fair comparison with the proposed 2DAL, we also use KMEM as the classifier.

We use the average $F1$ score over all different labels for performance evaluation, i.e., $F1 = \frac{2rp}{r+p}$ where p and r are precision and recall respectively. For the Scene data set, we use 241 (10%) images as the initial training set. In each iteration, 60 sample-label pairs are selected by the 2DAL. Note that, for 1DAL, it requests for annotation on the basis of samples rather than sample-label pairs, so in each iteration, it selects 10 images for annotating all the six labels or equivalently 60 image-label pairs. The average F1 score is then computed over all the remaining unlabeled data. In Figure 4(a), we show the performance of the three strategies over the total number of the selected sample-label pairs. The proposed 2DAL has the best performance over all iterations. With the number of selected pairs increasing, the improvement becomes more and more significant. Table 2 compares the $F1$ scores after 100 active learning iterations over all the six scene categories. The proposed 2DAL outperforms the other strategies on all the categories. In particular, the improvement is obvious on "Urban". Such an improvement is obtained by considering its significant correlations with other categories (see Figure 3 for an illustration of these label correlations) during the active learning procedure. It confirms 2DAL can obviously improve the classification performance.

4.2. Gene data set

The second data set is the Yeast data set [11] which consists of micro-array expression data and phylogenetic profiles with 2,417 genes and each gene in the set belongs to one or more of 14 different functional classes. As for multi-labeled gene data set, there are $33,838$ sample-label pairs in the sets. Each sample in this data set is annotated by 11 positive labels at most. The detailed description about this biological data set can be found in [6].

In the experiment, 242 (10%) genes with their labels are used as the initial training set. In each iteration, 140 sample-label pairs are selected. Similar to section 4.1, the 1DAL selects 14 samples for annotating all their labels. That's equivalent to 140 sample-label pairs. Figure 4(b) illustrates the performance of the three strategies on this data set.

From the above two experiments, we have observed:

1 When given a fixed number of annotations, 2DAL outperforms 1DAL over all the active learning iterations. This is because the former considers both sample and label uncertainty for selecting sample-label pair, while 1DAL only considers the sample uncertainty. Therefore, the informative label correlations associated with each sample can help to reduce the expensive human labors needed to construct the labeled pool.

2 The proposed 2DAL gives good performance on diverse data sets, ranging from natural scenes to gene images. This is an important character of a good algorithm to be used in real-world applications.

5. Conclusion

In this paper, we proposed an efficient two dimensional active learning (2DAL) strategy for multi-labeled image classification. This 2DAL strategy selects the sample-label pairs to annotate, along both the sample and label dimensions. In contrast to traditional one-dimensional binary active learning algorithms, 2DAL only needs to annotate a sub set of labels associated with a certain sample, thus much more efficient. Furthermore, we showed that the traditional active learning formulation is a special case of 2DAL when there is only one lable. Extensive experiments on two widely used data sets have shown that for a given number of required annotations, the proposed 2DAL strategy outperforms other state-of-the-art sample selection strategies.

Appendix

Here we give the proof of Lemma 1.

Proof. Since the selected y_s can take on two values $\{0,1\}$, there are two possible posterior distributions for the unlabeled y_i, i.e., $P\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)$ and $P\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)$. If $y_s=1$ holds, the Bayesian classification error is [7]:

$$\begin{aligned}&\mathcal{E}\left(y_i|y_s=1;y_{L(x)},x\right)=\min\{P\left(y_i=1|y_s=1;y_{L(x)},x\right)\\&,P\left(y_i=0|y_s=1;y_{L(x)},x\right)\}\end{aligned} \tag{21}$$

Given the inequality $\frac{1}{2}H(p)-\epsilon\leq\min\{p,1-p\}\leq\frac{1}{2}H(p),\epsilon=\frac{1}{2}\log\frac{5}{4}$ (see figure 5), we have

$$\begin{aligned}&\tfrac{1}{2}H\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)-\epsilon\leq\mathcal{E}\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&\leq\tfrac{1}{2}H\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\end{aligned} \tag{22}$$

Similarly, if $y_s=0$ holds,

$$\begin{aligned}&\tfrac{1}{2}H\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)-\epsilon\leq\mathcal{E}\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&\leq\tfrac{1}{2}H\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right).\end{aligned} \tag{23}$$

Therefore, the Bayesian classification error bound given the selected y_s can be computed as:

$$\begin{aligned}&\mathcal{E}\left(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&=P\left(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\mathcal{E}\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&+P\left(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\mathcal{E}\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&\leq\tfrac{1}{2}P\left(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)H\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&+\tfrac{1}{2}P\left(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)H\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&=\tfrac{1}{2}H\left(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\end{aligned} \tag{24}$$

The last equality follows the definition of conditional entropy. And similarly

$$\begin{aligned}&\mathcal{E}\left(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&=P\left(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\mathcal{E}\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&+P\left(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\mathcal{E}\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\\&\geq\tfrac{1}{2}P\left(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\left\{H\left(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)-2\epsilon\right\}\\&+\tfrac{1}{2}P\left(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x}\right)\left\{H\left(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)-2\epsilon\right\}\\&=\tfrac{1}{2}H\left(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)-\epsilon\end{aligned} \tag{25}$$

□

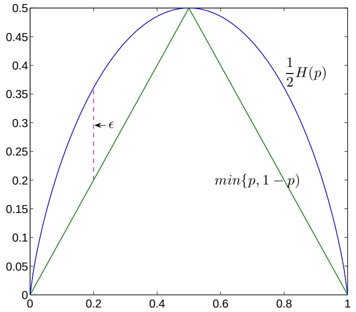


Figure 5. Illustration of the inequality $\frac{1}{2}H(p)-\epsilon\leq\min\{p,1-p\}\leq\frac{1}{2}H(p),\epsilon=\frac{1}{2}\log\frac{5}{4}$

References

[1] S. Benson, L. C. McInnes, J. Moré, T. Munson, and J. Sarich. TAO user manual (revision 1.9). Technical Report ANL/MCS-TM-242, Mathematics and Computer Science Division, Argonne National Laboratory, 2007. http://www.mcs.anl.gov/tao. 6

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9), 2004. 1, 6

[3] K. Brinker. On active learning in multi-label classification. *"From Data and Information Analysis to Knowledge Engineering" of Book Series "Studies in Classification, Data Analysis, and Knowledge Organization", Springer*, 2006. 1, 2

[4] E. Y. Chang, S. Tong, K. Goh, and C. Chang. Support vector machine concept-dependent active learning for image retrieval. *IEEE Transaction on Multimedia*, 2005. 1

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society (Series B)*, 39(1), 1977. 2, 6

[6] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Proc. of NIPS*, 2002. 1, 7

[7] M. E. Hellman and J. Raviv. Probability of error , equivocation, and the chernoff bound. *IEEE Transaction on Information Theory*, 1970. 3, 8

[8] S. C. H. Hoi and M. R. Lyu. A semi-supervised active learning framework for image retrieval. In *Proc. of IEEE CVPR*, 2005. 1

[9] F. Jing, M. Li, and H.-J. Zhang. Entropy-based active learning with support vector machine for content-based image retrieval. In *Proc. of IEEE Conference on Multimedia and Expo(ICME)*, 2004. 4

[10] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrel. Active learning with gaussian processes for object categorization. In *Proc. of IEEE ICCV*, 2007. 1, 4

[11] X. Li, L. Wang, and E. Sung. Multi-label svm active learning for image classification. In *Proc. of ICIP*, 2004. 1, 2, 7

[12] G.-J. Qi, X.-S. Hua, Y. Rui, J.-H. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proc. of ACM Conference on Multimedia (ACM Multimedia)*, 2007. 1

[13] N. Roy and A. McCallum. Toward optimal active learning through sampling esitmation of error reduction. In *Proc. of ICML*, 2001. 4

[14] S. Tong and E. Y. Chang. Support vector machine active learning for image retrieval. In *Proc. of ACM Conference on Multimedia (ACM Multimedia)*, 2001. 1, 4

[15] R. Yan, J. Yang, and A. Hauptmann. Automatically labeling data using multi-class active learning. In *Proc. of IEEE ICCV*, 2003. 1

[16] S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *Proc. of ACM SIGIR*, 2005. 1, 5