Active Learning with Multi-Label SVM Classification

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

Abstract

Multi-label classification, where each instance is assigned to multiple categories, is a prevalent problem in data analysis. However, annotations of multi-label instances are typically more time-consuming or expensive to obtain than annotations of single-label instances. Though active learning has been widely studied on reducing labeling effort for single-label problems, current research on multi-label active learning remains in a preliminary state. In this paper, we first propose two novel multi-label active learning strategies, a max-margin prediction uncertainty strategy and a label cardinality inconsistency strategy, and then integrate them into an adaptive framework of multi-label active learning. Our empirical results on multiple multi-label data sets demonstrate the efficacy of the proposed active instance selection strategies and the integrated active learning approach.

1 Introduction

Traditional multi-class classification problems assume that each instance is associated with a single label from a category set $\mathcal{Y}$, where $|\mathcal{Y}| > 2$. Multi-label classification generalizes multi-class classification by allowing each instance to be associated with multiple labels from $\mathcal{Y}$. In many real world data analysis problems, data objects can be assigned into multiple categories and hence produce multi-label classification problems. For example, an image for object categorization can be labeled as "desk" and "chair" simultaneously if it contains both objects. A news article talking about the effect of Olympic games on tourism industry might belong to multiple categories such as "sports", "economy", and "travel", since it may cover multiple topics.

Many approaches have been developed in the literature to address multi-label classification problems. One standard and simple solution for multi-label classification nevertheless is to generalize the "one-vs-all" scheme of multi-class classification. That is, one decomposes the multi-label problem into a set of binary classification problems, one for each class, and solves the multi-label classification problem by conducting standard binary classifications [Boutell *et al.*, 2004; Joachims, 1998; Lewis *et al.*, 2004a]. Regardless of the approach used, multi-label learning in general requires a sufficient amount of labeled data to recover high quality classification models. However, the labeling process of multi-label problems is much more expensive and time-consuming than single-label problems. In the single label case, a human annotator only needs to identify a single category to complete an instance label, whereas in the multi-label case, the annotator must consider every possible label for each instance, even if the positive labels are sparse. Active learning, which aims on conducting selective instance labeling and reducing the labeling effort of training good prediction models, is therefore particularly important for multi-label classification.

Despite the importance of the problem, current research on active learning for multi-label classification remains in a preliminary state. The majority of active learning study in the literature has centered on single-label classification problems, especially binary classification problems [Settles, 2012]. The active learning strategies developed for single-label classifications however mostly are not directly well applicable in multi-label cases, since instance selection decisions in multi-label cases should be based on all labels. The main challenge of multi-label active learning is to develop effective strategies to evaluate the unified informativeness of an unlabeled instance across all classes. Existing multi-label active learning works, such as [Brinker, 2006; Li *et al.*, 2004; Esuli and Sebastiani, 2009; Singh *et al.*, 2009; Yang *et al.*, 2009], measure the informativeness of an unlabeled instance by treating all labels in an independent way without considering the potential implicit label structure information across all classes.

In this paper, we propose two novel multi-label active learning strategies, a max-margin prediction uncertainty strategy and a label cardinality inconsistency strategy, which exploit the relative multi-label classification margin structure on each unlabeled instance and the statistical label cardinality information, respectively, to measure the unified informativeness of unlabeled instances. Moreover, we further investigate an adaptive integration framework of these two strategies by applying a novel approximate generalization error measure. Our empirical study on multiple multi-label classification data sets demonstrates the efficacy of the proposed multi-label active learning strategies and the integrated adaptive active learning approach.

2 Related Work

The aim of active learning is to reduce labeling effort and cost required for training a high quality prediction model. Given a large pool of unlabeled instances, an active learner iteratively selects most informative instances from the pool to query an oracle (e.g., a human annotator) for labels. Most active learning studies in the literature have focused on single-label classification problems. One most commonly used active learning strategy is *uncertainty sampling*, where the active learner selects the instance that is most uncertain to label for the current trained classification model. Though uncertainty sampling methods remain myopic without measuring the future predictive informativeness of the candidate instance on the large amount of unlabeled data, they are computationally efficient and have demonstrated good empirical performance [Lewis and Gale, 1994; Luo *et al.*, 2005; Culotta and McCallum, 2005; Settles and Craven, 2008]. Some more sophisticated non-myopic active learning methods exploit unlabeled data to minimize an *approximation of the generalization error* [Guo and Greiner, 2007; Guo and Schuurmans, 2007; Roy and McCallum, 2001; Yan *et al.*, 2003; Zhu *et al.*, 2003]. Such methods however are usually computationally expensive because they require a new prediction model to be retrained for each candidate query.

Active learning for multi-label classification however is still in a preliminary state. Most multi-label active learning methods decompose multi-label classification into a set of binary classification problems and make instance selection decisions by exploiting the binary classifiers independently without considering the label structure information of an instance revealed across all classes. [Brinker, 2006] uses a simple extension of the uncertainty sampling strategy. It decomposes the multi-label classification problem into several binary ones using the one-vs-all scheme, and selects the instance that minimizes the smallest SVM margin among all binary classifiers. [Singh *et al.*, 2009] simply takes the average of the uncertainty scores from all SVM binary classifiers as the instance selection measure. In [Li *et al.*, 2004], an SVM active learning method was proposed for multi-label image classification. It determines the predicted labels of an unlabeled instance using binary SVM classifiers and make instance selection decision by using Max Loss (ML) and Mean Max Loss (MML) strategies to count prediction losses of all binary classifiers. [Yang *et al.*, 2009] presents a strategy called maximum loss reduction with maximal confidence (MMC). It uses a multi-class logistic regression to predict the number of labels for an unlabeled instance and then computes the MMC measure by summing up losses from SVM classifiers on all labels. Different from these methods above, [Esuli and Sebastiani, 2009] exploits a multi-label boosting classification method and tests a number of strategies that conduct instance selections by combining measures from each class in an unequally weighted way. In addition, some other multi-label active learners consider selecting both instance and labels for annotations. For example, [Qi *et al.*, 2009] develops a two-dimensional active learning algorithm that selects sample-label pairs to minimize the Bayesian classification error bound. [Vijayanarasimhan and Grauman, 2009] develops a multi-label multiple-instance active learning approach that selects both an image example and a level of annotation to request. In this work, we nevertheless focus on the general problem of instance selection, and develop two novel multi-label uncertainty sampling strategies and an approximate generalization error measure to efficiently select the most informative instance.

3 Multi-label SVM Classification

Transforming a multi-label classification problem into a set of independent binary classification problems via the "one-vs-all" scheme is a conceptually simple and computationally efficient solution for multi-label classification. In this work, we conduct multi-label learning under such a mechanism by using standard support vector machines (SVMs) for the binary classification problems associated with each class.

Given a labeled multi-label training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i$ is the input feature vector for the i-th instance, and its label vector $\mathbf{y}_i$ is a $\{+1, -1\}$-valued vector with length K such as $K = |\mathcal{Y}|$. If $\mathbf{y}_{ik} = 1$, it indicates that the instance $\mathbf{x}_i$ is assigned into the k-th class; otherwise, the instance does not belong to the k-th class. For the k-th class ($k = 1, \cdots, K$), the binary SVM training is a standard quadratic optimization problem:

$$\min_{\mathbf{w}_k, b_k, \{\xi_{ik}\}} \quad \frac{1}{2}\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{N} \xi_{ik} \tag{1}$$

$$\text{subject to} \quad \mathbf{y}_{ik}(\mathbf{w}_k^T \mathbf{x_i} + b_k) \geq 1 - \xi_{ik},\ \xi_{ik} \geq 0,\ \forall i;$$

where $\{\xi_{ik}\}$ are the slack variables and C is the trade-off parameter. It maximizes the soft class separation margin. The model parameters $\mathbf{w}_k$ and b_k returned by this binary learning problem define a binary classifier associated with the k-th class: $f_k(\mathbf{x}_i) = \mathbf{w}_k^T \mathbf{x}_i + b_k$. The set of binary classifiers from all classes can be used independently to predict the label vector $\hat{\mathbf{y}}$ for an unlabeled instance $\hat{\mathbf{x}}$. The k-th component of the label vector $\hat{\mathbf{y}}_k$ has value 1 if $f_k(\hat{\mathbf{x}}) > 0$, and has value -1 otherwise. The absolute value $|f_k(\hat{\mathbf{x}})|$ can be viewed as a *confidence* value for its prediction $\hat{\mathbf{y}}_k$ on instance $\hat{\mathbf{x}}$.

4 Max-Margin Multi-label Active Learning

In this paper, we consider pool-based active learning which appears to be the most popular scenario for applied research in active learning. Assume we have a small set of labeled multi-label instances $\mathcal{L} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_\ell}$, but a large pool of unlabeled instances $\mathcal{U} = \{(\mathbf{x}_i)\}_{i=1}^{N_u}$. Same as above, the label vector $\mathbf{y}_i$ is a $\{+1, -1\}$-valued vector with length K. An active learner will iteratively select the most informative instance from the unlabeled pool $\mathcal{U}$ to label, then move it to the labeled set $\mathcal{L}$ and retrain the classification model on the augmented $\mathcal{L}$. We aim to design multi-label active learning strategies for learning a good multi-label SVM classification model with fewer labeled instances and hence lower labeling cost. Below we will first present two novel multi-label uncertainty sampling strategies from the perspectives of label prediction and label dimension statistics respectively, and then present an adaptive integration of these two strategies under a novel approximate generalization error measure.

4.1 Max-Margin Uncertainty Sampling

Uncertainty sampling is one of the simplest and most effective active learning strategies used for single-label classification. The central idea of this strategy is that the active learner should query the instance which the current classifier is most uncertain about. For binary SVM classifiers, the most uncertain instance can be interpreted as the one closest to the classification boundary [Campbell *et al.*, 2000]. As we reviewed in previous section, many multi-label active learning methods simply extend this binary uncertainty concept into the multi-label learning scenarios by integrating the binary uncertainty measures associated with each individual class in independent manners, such as taking the minimum over all classes [Brinker, 2006], and taking the average over all classes [Singh *et al.*, 2009; Yang *et al.*, 2009].

However, though multi-label classification can be conducted by training a set of independent binary classifiers, the prediction values produced by the multiple binary classifiers over the same instance are not irrelevant to each other. Note the training processes of multi-label classification and multi-class classification via the "one-vs-all" scheme are exactly the same. However, for a new instance $\mathbf{x}$, multi-class classification determines its single positive label by comparing the prediction values of all binary classifiers, such as $\mathbf{y}_{k^*} = 1$ for $k^* = \arg\max_k f_k(\mathbf{x})$. [Rifkin and Klautau, 2004] shows such a simple "one-vs-all" multi-class classifier is as accurate as any other multi-class approach, assuming the underlying binary classifiers are well-tuned regularized classifiers such as SVMs. This suggests the prediction values of binary SVM classifiers trained using the "one-vs-all" scheme for multi-label classification are directly comparable as well.

Moreover, inspired by ranking-loss based multi-label classification methods [Crammer and Singer, 2003; Guo and Schuurmans, 2011], we observe that multi-label prediction is really about the overall separation of the group of positive labels from the group of negative labels. We thus propose to use a global separation margin between the group of positive label prediction values and the group of negative label prediction values to model the prediction uncertainty of an instance under the current multi-label SVM classifiers. Specifically, given the set of binary SVM classifiers $f_1, \cdots, f_k$, the predicted label vector $\hat{\mathbf{y}}_i$ of an unlabeled instance $\mathbf{x}_i$ can be determined by the sign of the prediction values such as $\hat{\mathbf{y}}_{ik} = sign(f_k(\mathbf{x}_i))$. Let $\hat{\mathbf{y}}_i^+$ denote the set of predicted positive labels and $\hat{\mathbf{y}}_i^-$ denote the set of predicted negative labels, the separation margin over instance $\mathbf{x}_i$ can then be defined as

$$\begin{aligned} sep\_margin(\mathbf{x}_i) &= \min_{k \in \hat{\mathbf{y}}_i^+} f_k(\mathbf{x}_i) - \max_{s \in \hat{\mathbf{y}}_i^-} f_s(\mathbf{x}_i) \\ &= \min_{k \in \hat{\mathbf{y}}_i^+} |f_k(\mathbf{x}_i)| + \min_{s \in \hat{\mathbf{y}}_i^-} |f_s(\mathbf{x}_i)| \end{aligned} \quad (2)$$

Intuitively, a good multi-label classification model should *maximize* such separation margins over all instances to make sure the positive labels and the negative labels are well separated. The instance that has the smallest separation margin should be the most uncertain instance under the current classification model. Thus we define a novel *global* multi-label uncertainty measure as the inverse separation margin

$$u(\mathbf{x}) = \frac{1}{sep\_margin(\mathbf{x})} \quad (3)$$

We call this measure a *max-margin prediction uncertainty* measure since it aims to reduce the prediction uncertainty and increase the separation margins of all instances.

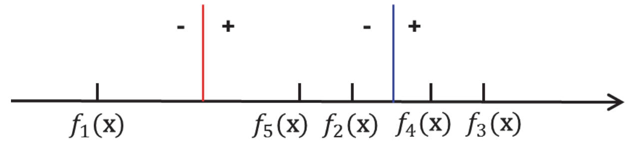


Figure 1: The ordered prediction values over instance $\mathbf{x}$ by binary classifiers across five classes. The red line marks the predicted separation line between positive and negative labels. The blue line marks the true separation line between positive and negative labels.

4.2 Label Cardinality Inconsistency

. The separation margin we defined above is computed based on the predicted positive labels and negative labels. However, when there are mistakes in label prediction, the predicted separation margin of an instance may not correctly reveal its prediction uncertainty property. Figure 1 demonstrates such a toy example with five classes, where the predicted separation margin of the example instance is large, but the prediction mistakes show the instance is in fact very uncertain to predict. Though it is impossible to identify exact prediction mistakes on unlabeled instances, we observe that the number of predicted positive labels can shed some useful information over the possible prediction mistakes and overall prediction uncertainty of an unlabeled instance according to the statistical dimension of positive labels.

The labeled and unlabeled instances are all drawn from the same underlying distribution, thus not only their input features, but also their output labels share common statistical properties. One observation we have is that the multi-label instances usually have similar number of positive labels. The average number of positive labels assigned to each instance in a multi-label data set is called its label cardinality [Tsoumakas and Katakis, 2007]. Thus the number of predicted positive labels of an unlabeled instance is expected to be consistent with the label cardinality computed on the labeled data. Based on this observation, we introduce a novel active selection strategy called *label cardinality inconsistency* to measure the prediction uncertainty over an unlabeled instance from the label dimension perspective. For an unlabeled instance $\mathbf{x}_i$, this inconsistency measure is defined as the Euclidean distance between the number of predicted positive labels and the label cardinality of the current labeled data:

$$c(\mathbf{x}_i) = \left\| \sum_{k=1}^{K} I_{[\hat{\mathbf{y}}_{ik} > 0]} - \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} \sum_{k=1}^{K} I_{[\mathbf{y}_{jk} > 0]} \right\|_2 \quad (4)$$

where $I_{[\cdot]}$ is an indicator function and it has value 1 when the given condition is true, 0 otherwise. Though very simple, our

Algorithm 1 Adaptive Active Learning Procedure

Input: labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$, parameter set B.
repeat
  Train multi-label SVM classifiers F^0 on $\mathcal{L}$.
  for each $\mathbf{x}_i \in \mathcal{U}$ **do**
    Compute $u(\mathbf{x}_i)$ and $c(\mathbf{x}_i)$.
  end for
  for each $\beta \in B$ **do**
    Mark a candidate instance $\mathbf{x} = \arg\max_{\mathbf{x}\in\mathcal{U}} q(\mathbf{x}, \beta)$.
  end for
  Copy all marked candidate instances into a set $\mathcal{S}$.
  for each $\mathbf{x} \in \mathcal{S}$ **do**
    Produce $\hat{\mathbf{y}}$ using classifiers F^0.
    Retrain a new classifiers F on $(\mathbf{x}, \hat{\mathbf{y}}) \cup \mathcal{L}$.
    Compute $\varepsilon(\mathbf{x})$ using classifier F and Eq. (6).
  end for
  Select instance $\mathbf{x}^*$ from $\mathcal{S}$ using Eq. (7).
  Remove $\mathbf{x}^*$ from $\mathcal{U}$, query its label vector $\mathbf{y}^*$.
  Add $(\mathbf{x}^*, \mathbf{y}^*)$ into $\mathcal{L}$.
until enough instances are queried

empirical work presented later shows this instance selection measure works reasonably well, even better than a few other multi-label instance selection strategies.

To our knowledge, exploiting information from label dimension perspective for active instance selection has also been exploited in the MMC method but in an indirect way [Yang *et al.*, 2009]. It uses a multi-class logistic regression classifier to predict the number of positive labels, m, for an unlabeled instance and then compute the loss reduction measure based on this prediction.

4.3 An Adaptive Integration Approach

The two active learning strategies we proposed above can be complementary to each other in many cases. For the toy example given in Figure 1, if the label cardinality computed from the labeled data is 2, then the uncertainty of the example instance $\mathbf{x}$ can be captured by the *label cardinality inconsistency* measure thought it has a low uncertainty value under the *max-margin prediction uncertainty* measure. We thus propose to combine the strengths of the two measures by integrating them in a weighted form

$$q(\mathbf{x}, \beta) = u(\mathbf{x})^{\beta} \cdot c(\mathbf{x})^{1-\beta} \tag{5}$$

where $\beta \in [0, 1]$ is a trade-off parameter that balances the relative importance degrees of the two measures.

However, it is difficult to pick a fixed weight parameter β that works well in different phases of active learning process and different active learning scenarios, since the strength of each component measure may vary in different learning scenarios. It is important to conduct flexible selections over the β parameter to fit into different learning scenarios. A previous work [Donmez *et al.*, 2007] tackled dynamic active learning by making strategy switch across stages of active learning process, which nevertheless lacks a consistent selection criterion across iterations. To achieve a consistent but flexible application of the integration criterion (5), we propose to adaptively select the best integration parameter β^* in each iteration of the active learning. Though it is hard to make continuous β value selection, we propose to select β value from a prefixed set of discretely sampled values, e.g., $B = [0, 0.1, \cdots, 0.9, 1]$. For each β value in the given set B, we can select one instance from the unlabeled pool $\mathcal{U}$ using the integrated selection measure in (5). After collecting all selected instances (no more than $|B|$ instances) together into a set $\mathcal{S}$, we then select the best β value from B by selecting the most informative instance from the set $\mathcal{S}$.

To make β selection, we propose an approximate generalization error for refined instance selection from the pre-selected set $\mathcal{S}$, which measures the future prediction error if the candidate instance and its predicted labels were added to the labeled training set. Specifically, for each instance $\mathbf{x} \in \mathcal{S}$, we use the multi-label SVM classifier $F^0 = [f_1^0, \cdots, f_K^0]$ trained on the current labeled set $\mathcal{L}$ to predict its label vector $\hat{\mathbf{y}}$. Then we train a new multi-label SVM classifier $F = [f_1, \cdots, f_K]$ on the augmented labeled set $\mathcal{L} \cup (\mathbf{x}, \hat{\mathbf{y}})$. The *approximate generalization error* of this new classifier F induced by the candidate instance $\mathbf{x}$ is defined as

$$\varepsilon(\mathbf{x}) = \sum_{i=1}^{N_u} \max_{k\in\hat{\mathbf{y}}_i^+}[1 - f_k(\mathbf{x}_i)]_+ + \max_{s\in\hat{\mathbf{y}}_i^-}[1 + f_s(\mathbf{x}_i)]_+ \tag{6}$$

where $[a]_+ = \max(0, a)$, $\hat{\mathbf{y}}_i^+$ denotes the predicted positive labels of the unlabeled instance $\mathbf{x}_i$ by the classifier F, and $\hat{\mathbf{y}}_i^-$ denotes the predicted negative labels correspondingly. This error is simply the sum of the two hinge losses around the predicted separation margin on each unlabeled instance. Finally the instance selection on $\mathcal{S}$ can be conducted by

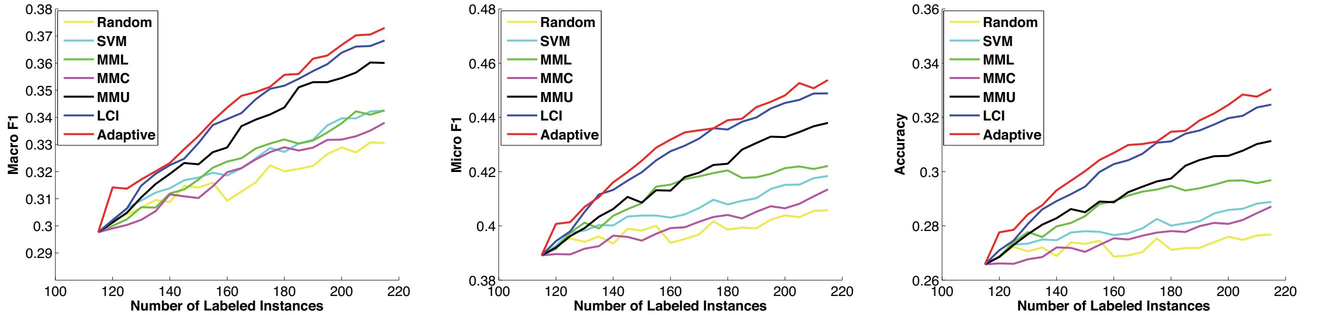
$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{S}} \ \varepsilon(\mathbf{x}) \tag{7}$$

It is natural to choose the unlabeled instance that would lead to the greatest reduction in future prediction error. However, it is computationally expensive to employ such a strategy directly because it requires retraining the multi-label classification model for each candidate instance in the unlabeled pool $\mathcal{U}$. Nevertheless, it is a suitable strategy to make refined instance selection from a pre-selected small set in our algorithm. The overall adaptive active learning procedure is described in Algorithm 1.

5 Experimental Results

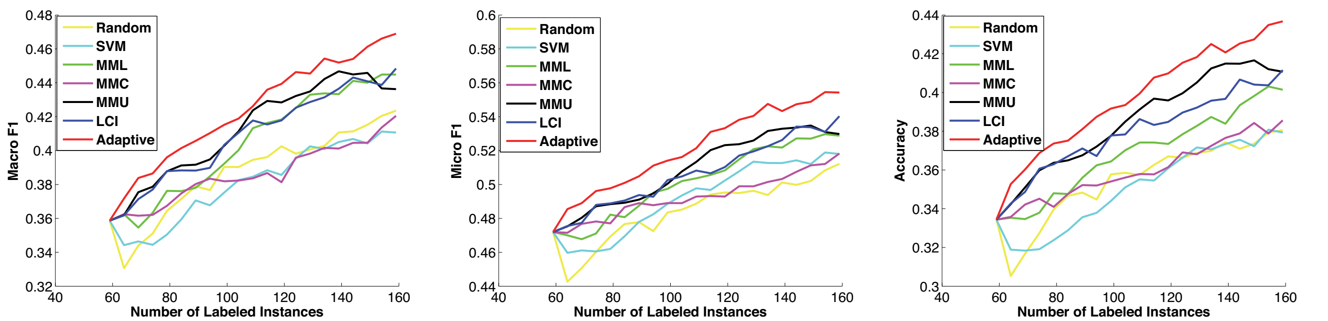
We evaluate our proposed multi-label active learning approach by conducting experiments on three image data sets, *Corel5K* [Duygulu *et al.*, 2002], *MSRC 23-class* [Shotton *et al.*, 2006], *MIR Flickr* [Huiskes and Lew, 2008], and a text data set, *RCV1-S2* [Lewis *et al.*, 2004b]. We compared the following approaches in our experiments:

- *Random*– the baseline using random instance selection.
- *SVM*– the baseline method that selects the most uncertain instance from all the uncertainty instances selected by the individual binary SVM classifiers, following the principle of the work [Brinker, 2006].
- *MML*– the method proposed in [Li *et al.*, 2004].
- *MMC*– the method proposed in [Yang *et al.*, 2009].



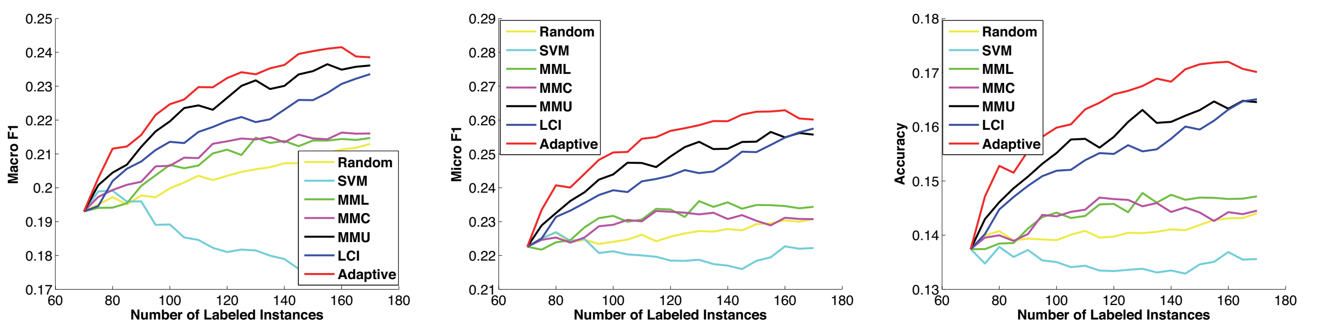
(a) Macro-F1 (b) Micro-F1 (c) Accuracy

Figure 2: The average results over 10 runs in terms of Macro-F1, Micro-F1 and Accuracy on the Corel5K subset with 15 classes.



(a) Macro-F1 (b) Micro-F1 (c) Accuracy

Figure 3: The average results over 10 runs in terms of Macro-F1, Micro-F1 and Accuracy on the MSRC 23-class data set.



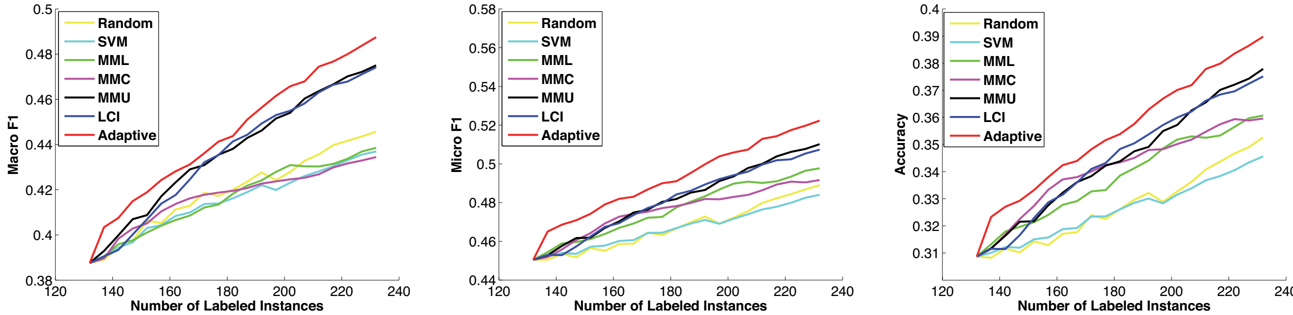
(a) Macro-F1 (b) Micro-F1 (c) Accuracy

Figure 4: The average results over 10 runs in terms of Macro-F1, Micro-F1 and Accuracy on the MIR Flickr data set.

- ***MMU***– the active learning method based on the *max-margin prediction uncertainty* sampling strategy we proposed in Section 4.1.
- ***LCI***– the active learning method based on the *label cardinality inconsistency* strategy in Section 4.2.
- ***Adaptive***– the adaptive active learning approach we developed in this paper.

All these methods use the multi-label SVM classification model for multi-label classification. We used a fixed trade-off parameter $C = 10$ in all the experiments.

Experimental Setting To conduct our active learning experiments, we sampled a 15-class subset of the Corel5K data with 2,160 images and a label cardinality value 2.4; a 15-class subset of the MIR Flickr data with 2,301 images and a label cardinality value 2.3. We used the entire MSRC 23-class data



(a) Macro-F1 (b) Micro-F1 (c) Accuracy

Figure 5: The average results over 10 runs in terms of Macro-F1, Micro-F1 and Accuracy on the RCV1-S2 subset.

set, which has 591 images over 23 classes and a label cardinality value 2.5. For these image classification tasks, we used GIST features [Oliva and Torralba, 2001] and SIFT [Lowe, 2004] features for image representation. For the RCV1-S2 text data, we sampled a 15-class subset with 2,657 documents in total and a label cardinality value 2.4.

For each active learning experiment, we first randomly partitioned the data into three parts: labeled set, unlabeled pool and test set, under the condition that at least one positive label appears for each class in the labeled set, and then ran each comparison approach independently to conduct active learning based on the same initial setting. The partition settings we used for the four data sets are given as below: Corel5K (115 labeled images; 1,496 unlabeled images; 690 test images); MSRC (59 labeled images; 354 unlabeled images; 177 test images); MIR Flickr (70 labeled images; 1,582 unlabeled images; 708 test images); and RCV1-S2 (132 labeled images; 1,727 unlabeled images; 797 test images). For each active learning approach, we ran it for 100 iterations, and queried 100 instances in total. In each iteration, after querying the label of the selected instance, we retrained the multi-label SVM classifier on the increased labeled set, and evaluated its performance on the test set in terms of three performance measures: macro-F1, micro-F1 and accuracy. We repeated each experiment 10 times and reported the average results.

Results The experimental results on the four data sets are reported in Figure 2 – Figure 5. We can see that the naive random sampling baseline, *Random*, obviously demonstrates inferior performance on all data sets, comparing to most other methods. The *SVM* method, which selects the most uncertain instance based on independent selections made by individual binary classifiers, demonstrates poor performance as well. Especially, on the *MIR Flickr* data set, the classification performance produced by *SVM* even degrades with more instances (possibly outliers) being labeled. The two specialized multi-label active learning methods, *MMC* and *MML*, demonstrate superior performance over the two baselines in many cases. *MMC*, originally introduced for document classification, outperforms both *Random* and *SVM* on the multi-label text classification data, RCV1-S2, in terms of micro-F1 and accuracy, and on the image data *MIR Flickr*, in terms of all three evaluation measures. *MML*, originally developed for image classification, outperforms the previous three methods, *Random, SVM* and *MMC* on the image data sets, Corel5K and MSRC 23-class, but produces similar performance as *MMC* on MIR Flickr and RCV1-S2. Nevertheless, the performance gains achieved by *MMC* and *MML* are small and inconsistent across different data sets. The proposed two novel active learning methods, *MMU* and *LCI*, on the other hand, demonstrate clear and consistent advantages over the previous four methods, *Random, SVM, MML* and *MMC*. *MMU* outperforms the previous four baseline methods on all data sets, and outperforms *LCI* on two data sets, MSRC 23-class and MIR Flickr. The simple label cardinality inconsistency based method, *LCI*, outperforms all four baseline methods on three data sets, *Corel5K, MIR Flickr*, and *RCV1-S2*. On *RCV1-S2*, *LCI* produces similar performance as *MMU*, and on *Corel5K*, *LCI* even outperforms *MMU*. This suggests the uncertainty knowledge based on simple label dimension statistics is very useful. The proposed *Adaptive* active learning method effectively combines the strengths of *MMU* and *LCI*, and has demonstrated outstanding superior performance comparing to all the other six comparison methods on all four data sets across all three evaluation measures.

6 Conclusions

In this paper, we proposed two novel multi-label active learning strategies, a max-margin prediction uncertainty strategy that exploits the relative multi-label classification margin structure of each unlabeled instance, and a label cardinality inconsistency strategy that exploits the statistical label cardinality information of the labeled data, to measure the unified informativeness of an unlabeled instance across multiple labels. Moreover, we further proposed to integrate the strengths of these two strategies using an adaptive integration framework, which relies on a novel approximate generalization error for refined instance selection. Our empirical study on multiple multi-label classification data sets from different application areas shows that the proposed multi-label active learning strategies, especially the integrated active learning approach, greatly outperform a number of multi-label active learning methods developed in the literature.

References

[Boutell *et al.*, 2004] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

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

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

[Crammer and Singer, 2003] K. Crammer and Y. Singer. A family of additive online algorithms for category ranking. *JMLR*, 3:1025–1058, 2003.

[Culotta and McCallum, 2005] A. Culotta and A. McCallum. Reducing labeling effort for structured prediction tasks. In *Proc. of AAAI*, 2005.

[Donmez *et al.*, 2007] Pinar Donmez, Jaime G. Carbonell, and Paul N. Bennett. Dual strategy active learning. In *Proc. of ECML*, 2007.

[Duygulu *et al.*, 2002] P. Duygulu, K. Barnard, J de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, 2002.

[Esuli and Sebastiani, 2009] A. Esuli and F. Sebastiani. Active learning strategies for multi-label text classification. In *Proc. of ECIR*, 2009.

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

[Guo and Schuurmans, 2007] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *Proc. of NIPS*, 2007.

[Guo and Schuurmans, 2011] Y. Guo and D. Schuurmans. Adaptive large margin training for multilabel classification. In *Proc. of AAAI*, 2011.

[Huiskes and Lew, 2008] M. Huiskes and M. Lew. The MIR flickr retrieval evaluation. In *Proc. of ACM international conference on Multimedia information retrieval*, 2008.

[Joachims, 1998] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proc. of ECML*, 1998.

[Lewis and Gale, 1994] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proc. of Annual Inter. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[Lewis *et al.*, 2004a] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397, 2004.

[Lewis *et al.*, 2004b] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397, 2004.

[Li *et al.*, 2004] X. Li, L. Wang, and E. Sung. Multilabel SVM active learning for image classification. In *Proc. of ICIP*, 2004.

[Lowe, 2004] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Luo *et al.*, 2005] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. Active learning to recognize multiple types of plankton. *JMLR*, 6:589–613, 2005.

[Oliva and Torralba, 2001] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[Qi *et al.*, 2009] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *TPAMI*, 31(10):1880 –1897, 2009.

[Rifkin and Klautau, 2004] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *JMLR*, 5:101–141, 2004.

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

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

[Settles, 2012] B. Settles. *Active Learning*. Morgan & Claypool, 2012.

[Shotton *et al.*, 2006] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. of ECCV*, 2006.

[Singh *et al.*, 2009] M. Singh, E. Curran, and P. Cunningham. Active learning for multi-label image annotation. Technical report, University College Dublin, 2009.

[Tsoumakas and Katakis, 2007] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Inter. J. of Data Warehousing & Mining*, 3(3):1–13, 2007.

[Vijayanarasimhan and Grauman, 2009] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Proc. of CVPR*, 2009.

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

[Yang *et al.*, 2009] B. Yang, J. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *Proc. of ACM SIGKDD Inter. Conference on Knowledge Discovery and Data Mining*, 2009.

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