Active Query Driven by Uncertainty and Diversity for Incremental Multi-Label Learning *

Sheng-Jun Huang and Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{*huangsj, zhouzh*}@*lamda.nju.edu.cn*

***Abstract*—In multi-label learning, it is rather expensive to label instances since they are simultaneously associated with multiple labels. Therefore, active learning, which reduces the labeling cost by actively querying the labels of the most valuable data, becomes particularly important for multi-label learning. A strong multi-label active learning algorithm usually consists of two crucial elements: a reasonable criterion to evaluate the gain of queried label, and an effective classification model, based on whose prediction the criterion can be accurately computed. In this paper, we first introduce an effective multi-label classification model by combining label ranking with threshold learning, which is incrementally trained to avoid retraining from scratch after every query. Based on this model, we then propose to exploit both uncertainty and diversity in the instance space as well as the label space, and actively query the instance-label pairs which can improve the classification model most. Experimental results demonstrate the superiority of the proposed approach to state-of-the-art methods.**

***Keywords*-active learning; multi-label learning; uncertainty; diversity**

I. Introduction

In many applications, we have plenty of unlabeled data but few labeled data. While labeling is usually expensive since it requires the participation of human experts, training an accurate model with as few labeled data as possible becomes a challenge of great significance. Active learning, which reduces the labeling cost by actively selecting the most valuable data to query their labels, is a leading approach to this goal [23]. The key task in active learning is to design a selection criterion such that queried labels can improve the classification model most. During the past years, many active selection criteria have been proposed. For example, *uncertainty* measures the confidence of the current model on classifying an instance [26], *diversity* measures how different an instance is from the labeled data [2], *density* measures the representativeness of an instance to the whole data set [18], and so on. There are also some other approaches trying to consider different criteria simultaneously [6], [13].

In traditional supervised classification problems, one instance is assumed to be associated with only one label. However, in many real world applications, an object can have multiple labels simultaneously. Multi-label learning is a framework dealing with such objects [32]. To label the multi-label examples, each of the multiple labels should be decided whether a proper one for an instance. Obviously, the labeling cost is even higher than that of single label learning, and thus active learning under the multi-label setting has attracted more and more attention.

In multi-label active learning, the main efforts focus on exploiting uncertainty, leaving the other active selection criteria rarely considered. Different algorithms are designed to evaluate the uncertainty of unlabeled data. A commonness of them is that they usually evaluate the active selection criterion based on the predicted labels of instances. Thus an effective classifier which can accurately predict the labels for the unlabeled data is crucial for a successful active learning algorithm. Most existing methods decompose the multi-label task into a series of binary classification problems and learn each label independently. However, the prediction values on different labels may not be comparable, and the number of positive labels for each instance is unknown, thus it is still a challenge to decide the threshold for separating positive and negative labels of an instance given the prediction values on each label. To address this challenge, some ad-hoc efforts have been attempted. For example, simply taking the sign of predictions as labels [16], normalizing the predictions on different labels [24], and predicting the number of positive labels via an extra regression model [30].

In this paper, we propose to exploit both uncertainty and diversity in the instance space and label space with an incremental multi-label classification model. First, along with a label ranking algorithm which optimizes the approximated ranking loss to rank positive labels before negative ones, we also employ a dummy label for each instance, and train the model to rank the dummy label between positive and negative labels. Since the dummy label threshold is learned specifically for each instance along with the ranking model, it is expected to provide an accurate separation of positive and negative labels. Based on this model, we then integrate uncertainty with diversity as a new active learning criterion, and select the most valuable instance-label pairs to query. Specifically, in the instance space, we simultaneously evaluate the uncertainty with label cardinality inconsistency and

*This work was partially supported by NSFC (61073097, 61273301) and JiangsuSF (BK2011566)

the diversity with the number of labels not queried; while in the labels space, the distance from a label to the thresholding dummy label is employed to evaluate the uncertainty. After each query, our multi-label model is incrementally updated based on only the newly added labeled data, avoiding the retraining from scratch.

The rest of this paper is organized as follows. Section II reviews related work. In Section III, the multi-label classification model as well as the active selection strategy are presented. Section IV presents the experiments, followed by the conclusion in Section V.

II. Related Work

The multi-label classification model used in this paper is extended from an online multi-class ranking model [29]. The method in [29] is designed for single label setting, and thus can not decide how many labels should be selected as positive from the ranked label list. In this paper, we employ a dummy label to automatically learn a threshold for separating positive and negative labels, which is previously used in [12] for multi-instance multi-label learning. Fürnkranz et al. propose a calibrated label ranking approach for multi-label classification in [11], where a mechanism similar to our dummy label is used for separating positive and negative labels. However, to the best of our knowledge, label ranking with threshold learning has not been exploited in multi-label active learning.

Though the main efforts on active learning focused on the single-label setting, multi-label active learning has attracted more and more attention due to its importance [14]. Most existing approaches are based on the binary relevance model, which decomposes the multi-label task into a series of binary classification problems. For example, [17] trains a SVM for each label, and selects the instance which maximizes the reduction of expected loss to query its labels. [30] also trains a SVM for each label and aims to maximize the expected loss reduction, but uses an extra regression model to predict the number of labels for each instance. Besides, based on independently trained binary classifiers, the minimal, average and weighted summarization over the uncertainties measured on each label are taken as active selection criterion in [3], [25] and [9], respectively. In [16], label cardinality inconsistency is combined with the separation margin via a tradeoff parameter as a new criterion for active selection, where the labels of instances are directly determined by the sign of prediction values of individual binary SVMs.

While most multi-label active learning methods try to query the whole label vector of one instance at a time, Qi et al. propose a two dimensional approach to iteratively query whether a label is positive on a specific instance [20]. They derive a multi-labeled Bayesian error bound, and select the instance-label pairs which maximizes the expected error reduction to query.

There are also some other multi-label active learning approaches proposed for different settings, such as the method for multi-view data [33] and the batch model form algorithm proposed in [4].

III. The Algorithm

We first introduce an incremental multi-label model by extending the label ranking method proposed in [29] from single-label to the multi-label in subsection A, and then propose a new active learning strategy in subsection B to iteratively query whether a label is positive on an instance.

A. Multi-Label Ranking with Dummy Label

In [29], an online algorithm was proposed to optimize approximated ranking loss on single-label data, aiming to rank the positive label before negative ones for each instance. In single-label learning, we can easily determine the positive label for an instance by selecting the one with maximum prediction value. However, in multi-label learning where one instance can have more than one label, we do not know how many labels should be selected as positive from the label list ranked based on the predictions values. To overcome this difficulty, inspired by [11], [12], we introduce a dummy label to each instance, and train the model to rank the dummy label between positive and negative labels for thresholding.

We denote by $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ the training data with n examples, where each instance $\mathbf{x}_i$ is a d-dimensional feature vector. Assuming there are in all K possible labels, the label vector of $\mathbf{x}_i$ is denoted by $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{iK}]^\top$, where $y_{ik} = 1$ if instance $\mathbf{x}_i$ has the k-th label, otherwise $y_{ik} = -1$. The classification model first maps both the feature space and label space into a shared subspace, and then measures the relevance of a label to an instance by their similarity in this subspace. Formally, given an instance $\mathbf{x}$, we define the classification model on label k as

$$f_k(\mathbf{x}) = \mathbf{w}_k^\top W_0 \mathbf{x},$$

where W_0 is a $m \times d$ matrix which maps the original feature vectors to the subspace, and $\mathbf{w}_k$ is a m-dimensional vector that maps the k-th label into the subspace. Here d and m are the dimensionalities of the feature space and the subspace, respectively. Then the ranking error [28], [29] for the instance $\mathbf{x}$ and one of its positive labels y is defined as:

$$\epsilon(\mathbf{x}, y) = \sum_{i=1}^{R(\mathbf{x},y)} \frac{1}{i}, \tag{1}$$

where $R(\mathbf{x}, y)$ counts how many negative labels are ranked before label y. Obviously, the lower y is ranked, the larger $R(\mathbf{x}, y)$ is, and accordingly the ranking error $\epsilon(\mathbf{x}, y)$ becomes larger. By minimizing the ranking error on all positive labels of all training instances, the model is expected to rank positive labels before negative ones for an unseen instance.

Let $\bar{Y}$ denote the set of all negative labels of $\mathbf{x}$, as in [29], the ranking error $\epsilon(\mathbf{x}, y)$ can be approximated by a convex surrogate loss

$$\epsilon(\mathbf{x}, y) \approx \sum_{\bar{y} \in \bar{Y}} \epsilon(\mathbf{x}, y) \frac{|1 + f_{\bar{y}}(\mathbf{x}) - f_y(\mathbf{x})|_+}{R(\mathbf{x}, y)}, \qquad (2)$$

which is further spread into every negative label $\bar{y}$ in $\bar{Y}$:

$$\mathcal{L}(\mathbf{x}, y, \bar{y}) = \epsilon(\mathbf{x}, y)|1 + f_{\bar{y}}(\mathbf{x}) - f_y(\mathbf{x})|_+. \qquad (3)$$

This surrogate loss is minimized with stochastic gradient descent (SGD) [21]. At each iteration of SGD, we first randomly select an instance $\mathbf{x}$ and one of its positive labels y, and then iteratively sample negative labels $\bar{y}$ from $\bar{Y}$ until getting a violated label (i.e., $\bar{y}$ is ranked before y). Based on the number of samples needed to reach this violated label, $R(\mathbf{x}, y)$ can be estimated as $R(\mathbf{x}, y) = \sum_{\bar{y} \in \bar{Y}} I[f_{\bar{y}}(\mathbf{x}) > f_y(\mathbf{x}) - 1]$ according to [29]. After that, gradient descent is performed aiming to minimize $\mathcal{L}(\mathbf{x}, y, \bar{y})$. At last, the updated parameters, i.e., $\mathbf{w}_y$, $\mathbf{w}_{\bar{y}}$ and each column of W_0 are then normalized to have a ℓ^2-norm smaller than a specific constant C.

We then explain how the positive and negative labels are separated with the dummy label. We assume that every instances $\mathbf{x}_i$ has a dummy label, denoted by y_{i0}. At each iteration of SGD, the sampled label y can be a positive label of $\mathbf{x}$ or the dummy label y_0. As in [12], we construct the negative label set $\bar{Y}$ depending on the type of y. If y is the dummy label, then $\bar{Y}$ consists of all negative labels of instance $\mathbf{x}$, otherwise, $\bar{Y}$ contains all negative labels as well as the dummy label. After the training with such a mechanism, the dummy label will be ranked before all negative labels while positive labels will be ranked before both the dummy label and negative labels. So the dummy label provides a nature threshold to separate positive and negative labels. Given an unseen test instance, with the prediction values on each label, we can easily select the labels with larger predictions than that of the dummy label as positive labels.

B. Active Selection

In this subsection, we present the strategy of active selection based on the previously introduce multi-label classification model. We follow the setting in [20] to iteratively query if a label is positive on an instance. We denote by $\mathcal{D}$ the data set, and divide it into two parts: the labeled data $\mathcal{D}_l$ with N_l instances and unlabeled data $\mathcal{D}_u$ with N_u instances. Since we select instance-label pairs for querying, there will be some instances partially labeled. For convenience, such partially labeled instances are also taken as unlabeled data. In other words, D_l contains only the fully labeled instances, while $\mathcal{D}_u$ contains both the unlabeled and partially labeled instances. We also introduce $U(\mathbf{x})$ to denote the set of labels that have not been queried for instance $\mathbf{x}$. So the task in each iteration is to select an instance $\mathbf{x}^*$ from $\mathcal{D}_u$ and then select one label y^* from $U(\mathbf{x}^*)$ to query.

Uncertainty is an effective and mostly used criterion for active learning [16], [25], [26]. In this paper, we will try to exploit the uncertainty in both the instance space and label space, and combine it with diversity for active selection. In single-label learning, a classic implementation of uncertainty sampling is to query the instance closest to the decision boundary. This strategy can be easily extend to multi-label learning. For example, the average or minimal margin over all labels can be taken to measure the uncertainty of an instance [25], [9]. Recently, a simple uncertainty criterion, named as label cardinality inconsistency was proposed in [16]. It measures the inconsistency between the number of predicted positive labels of an instance and the average label cardinality on the fully labeled data, and can be formally defined as:

$$LCI(\mathbf{x}_i) = (\sum_{k=1}^{K} I[\hat{y}_{ik} > 0] - \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{k=1}^{K} I[y_{jk} > 0])^2,$$

where $\hat{y}_{ik}$ $(k = 1 \cdots K)$ is the predicted labels for instance $\mathbf{x}_i$, and $I[\cdot]$ is the indicator function. In [16], LCI was combined with margin with a tradeoff parameter to measure the uncertainty, and has been shown to be effective. However, with the increase of labeled data, the difference of LCI over different instances may get smaller and smaller, and thus it becomes more difficult to identify the most uncertain instance based on the LCI criterion. In this paper, we extend LCI to incorporate with diversity, and define a new criterion:

$$C_1(\mathbf{x}_i) = \frac{\left| \sum_{k=1}^{K} I[\hat{y}_{ik} > 0] - \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{k=1}^{K} I[y_{jk} > 0] \right|}{\max\{\xi, K - card(U(\mathbf{x}_i))\}}, \qquad (4)$$

where $card(\cdot)$ counts the number of elements in a set, and thus $K - card(U(\mathbf{x}_i))$ is the number of queried labels of $\mathbf{x}_i$. $\xi \in (0, 1)$ is a constant to avoid the zero divisor. The motivation here is that instances with less queried labels may contain more unknown information and should be preferentially queried. As we want to query the instance with maximum uncertainty and diversity, the instance $\mathbf{x}_*$ achieves maximum C_1 value is selected. Note in our experiments, we set $\xi = 0.5$ and randomly select one if multiple instances achieved the maximal C_1 value.

After selecting the instance $\mathbf{x}^*$, we need to decide which label to query. Since the dummy label stands for the separating threshold of the positive and negative labels, the uncertainty of a label y can be naturally measured by the distance from it to the dummy label, which is formally defined as:

$$C_2(\mathbf{x}^*, y) = |f_y(\mathbf{x}^*) - f_{y_0}(\mathbf{x}^*)| \qquad (5)$$

The pseudo code of the proposed algorithm, termed AUDI (Active learning based on Uncertainty and Diversity for

Algorithm 1 The AUDI algorithm

```
Input:
    data set $\mathcal{D}$
Initialize:
    Divide $\mathcal{D}$ to $\mathcal{D}_l$ and $\mathcal{D}_u$
    train a model $f$ on $\mathcal{D}_l$
Repeat:
    get predictions and labels for instances in $\mathcal{D}_u$ with $f$
    compute $C_1(\mathbf{x})$ for all $\mathbf{x} \in D_u$ as Eq. 4
    select the instance $\mathbf{x}^*$ with maximum $C_1$ value
    compute $C_2(\mathbf{x}^*, y)$ for all $y \in U(\mathbf{x}^*)$ as Eq. 5
    select the label $y^*$ with minimal $C_2$ value
    query if label $y^*$ is a positive one for instance $\mathbf{x}^*$
    remove $y^*$ from $U(\mathbf{x}^*)$
    move $\mathbf{x}^*$ from $\mathcal{D}_u$ to $\mathcal{D}_l$ if $|U(\mathbf{x}^*)| = 0$
    update the model $f$ with $\mathbf{x}^*$ and its labels
until the number of queries reached
```

Incremental multi-label learning), is presented in Algorithm 1. First, a subset of the data set $\mathcal{D}$ is randomly sampled to initialize the labeled data $\mathcal{D}_l$. Note in the experiments, to be fair, for all algorithms, $\mathcal{D}_l$ is initialized with the same set of fully labeled instances rather than instance-label pairs. After the initialization, we apply the algorithm introduced in the previous subsection on $\mathcal{D}_l$ to train a multi-label classification model. Then at each iteration of active learning, we select an instance-label pair $(\mathbf{x}^*, y^*)$ according to:

$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{D}_u} C_1(\mathbf{x})$$
$$y^* = \operatorname*{argmin}_{y \in U(\mathbf{x}^*)} C_2(\mathbf{x}^*, y)$$

and query if y^* is a positive label of $\mathbf{x}^*$. After that, y^* is removed from $U(\mathbf{x}^*)$ and $\mathbf{x}^*$ is moved from $\mathcal{D}_u$ to $\mathcal{D}_l$ if it is fully labeled. Since the multi-label classification model can be trained incrementally, we do not need to retrain the model from scratch, but only update f based on the newly labeled data. Note that adding a label to an instance may affect the rank of all labels on it, so f is updated on $\mathbf{x}^*$ and all of its labels, instead of only on the pair $(\mathbf{x}^*, y^*)$. This active querying and model updating process is repeated until enough data labeled.

IV. Experiments

A. Settings

In the experiments, the following six multi-label active learning approaches are compared:

- Random: the baseline which randomly selects instance-label pairs.
- 2DAL: the two dimensional active learning method proposed in [20], which selects instance-label pairs with the expected classification error reduction criterion.
- MML: the mean max loss strategy proposed in [17].
- MMC: the method proposed in [30], which uses the maximum loss reduction with maximal confidence as selection criterion.
- Adaptive: the adaptive method proposed in [16], which combines the max-margin prediction uncertainty and the label cardinality inconsistency as the criterion for active selection.
- AUDI: the method proposed in this paper.

Table I
Statistics on data sets used in the experiments

Data	# instance	# label	# feature	cardinality
Corel5K [7]	5000	374	499	3.52
Emotions [27]	593	6	72	1.87
Enron [15]	1702	53	1001	3.38
Genebase [5]	662	27	1185	1.25
Image [31]	2000	5	294	1.24
Medical [19]	978	45	1449	1.25
Reuters [22]	2000	7	243	1.15
Scene [1]	2407	6	294	1.07
Yeast [8]	2417	14	103	4.24

Experiments are performed on 9 data sets, most of which can be download at the web page of MULAN project[1]. Detailed characteristics of these data sets are summarized in Table I, including number of instances, number of labels, feature space dimensionality and label cardinality (LC), where LC counts the average number of labels per instance.

For each data set, we randomly divide it into two parts with equal size, take one part as test set and the other part as the unlabeled pool for active selection. The random data partition is repeated for 10 times, and average results over the 10 repeats are reported. At the very beginning of active learning, we randomly sample 5% instances from the unlabeled pool as initial labeled data. At each iteration of active learning, one instance or one instance-label pair is selected by the active learning methods based on their own strategy, and then added into the labeled data. After $2 \times m$ instance-label pairs queried, we train a classification model on the labeled data and evaluate its performance on the holdout test data. The querying process is stopped if all data are fully labeled or the number of queried instance-label pairs reaches 20000.

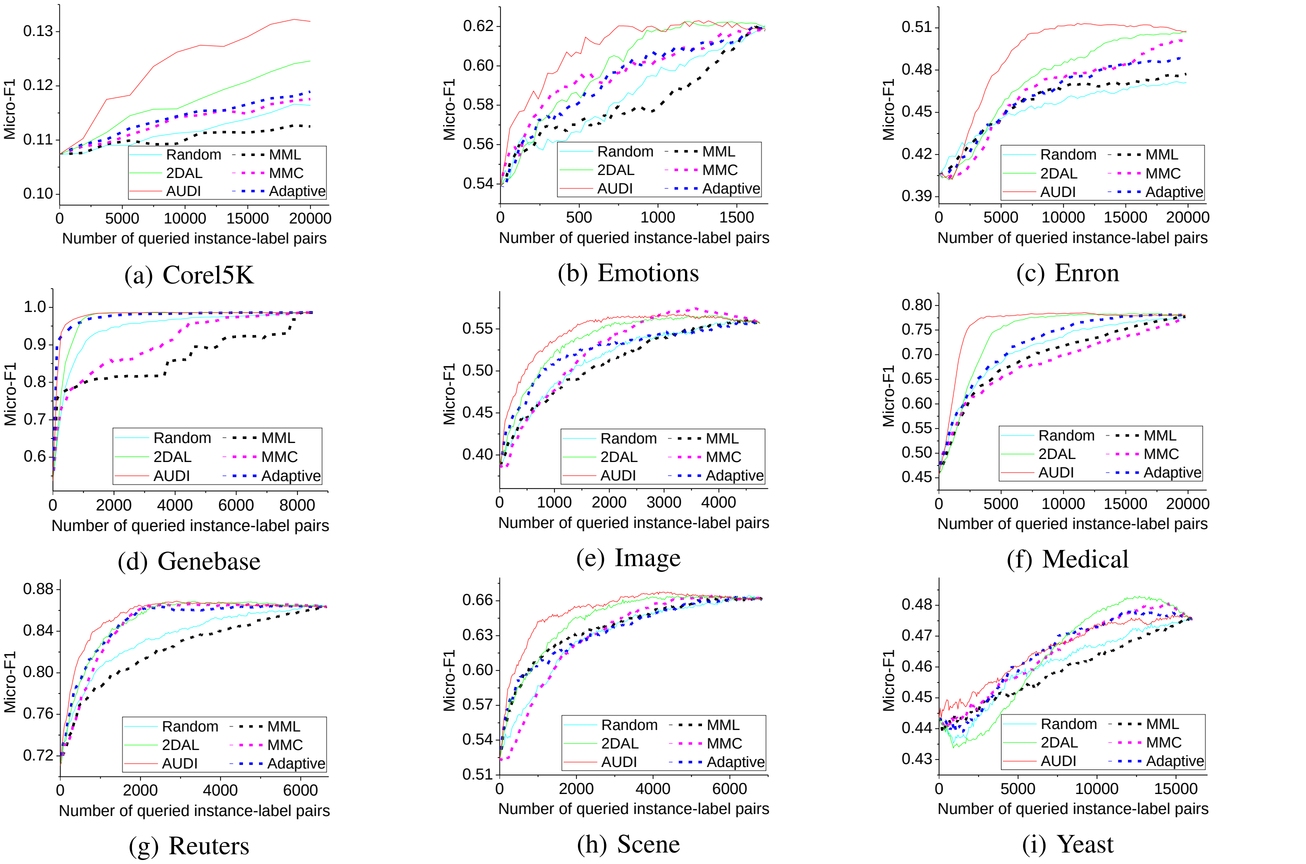
We evaluate the performances of compared approaches on micro-F1, which is commonly used in multi-label learning [16], [30]. Micro-F1 computes the F1 measure by considering predictions of all instances on all labels together. It is formally defined as Eqs. 6.

$$\text{micro-F1} = \frac{2\sum_{i=1}^{n}\sum_{k=1}^{K} I[y_{ik}=1] \cdot I[\hat{y}_{ik}=1]}{\sum_{i=1}^{n}\sum_{k=1}^{K}(I[y_{ik}=1] + I[\hat{y}_{ik}=1])}, \quad (6)$$

where $\hat{y}_{ik}$ denotes the predicted label of the i-th instance on the k-th label, and $I[\cdot]$ is the indicator function.

To be fair, we use one-versus-all linear SVM (implemented with LIBLINEAR [10]) as the classification model for evaluating all the compared approaches. For MMC, the

[1] http://mulan.sourceforge.net/datasets.html



(a) Corel5K

(b) Emotions

(c) Enron

(d) Genebase

(e) Image

(f) Medical

(g) Reuters

(h) Scene

(i) Yeast

Figure 1. Comparison results on *Micro-F1*.

regression model is also implemented with LIBLINEAR. For AUDI, we use constant step size for SGD and set $m = 100$ as default. The other parameters are selected via 5-folds cross validation on the initial labeled data. For the other approaches, parameters are determined in the same way if no values suggested in their literatures.

B. Comparison Results

We plot the curves of micro-F1 with the number of queried instance-label pairs increasing in Fig. 1. Note that three approaches: Random, 2DAL and AUDI, which query instance-label pairs, are plotted in solid line, while the other three methods which query the whole label vector of instances are plotted in dashed line.

Generally speaking, methods querying instance-label pairs are more effective than those query instances, which is consistent with the results in [20]. This phenomenon is probably because of that multiple labels may be correlated, and thus redundancy of information may exist among the multiple labels of the same instance. This also explains why random selection can be better than some active approaches on some data. Among the methods query instances only, Adaptive and MMC tend to be more effective than MML.

When comparing the proposed AUDI with other methods, no matter querying instance-label pairs or instances only, our method achieves the best performance in most cases. Especially on data sets with more labels, such as *Corel5K* and *Enron* data sets, the superiority of AUDI gets more significant. The only special case is on *yeast* data set, where AUDI achieves comparable performance with the best baseline.

V. Conclusion

An effective classification model along with a good selection criterion are the key factors of a successful active learning approach. In this paper, we propose a new multi-label active learning approach to iteratively query whether a label is positive on an instance. First, we present an incremental model for effective multi-label classification by incorporating label ranking with a threshold mechanism. Based on the model, we then propose to combine uncertainty and diversity as the criterion to select the most valuable instance-label pairs to query. The experimental results demonstrate the superiority of our algorithm. In the future, we will try to study other active query strategies based on the label ranking model.

References

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

[2] K. Brinker. Incorporating diversity in active learning with support vector machines. In *ICML*, pages 59–66, 2003.

[3] K. Brinker. On active learning in multi-label classification. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nrnberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, pages 206–213. Springer Berlin Heidelberg, 2006.

[4] S. Chakraborty, V. Balasubramanian, and S. Panchanathan. Optimal batch selection for active learning in multi-label classification. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1413–1416, 2011.

[5] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas. Protein classification with multiple algorithms. In *Proceedings of the 10th Panhelllenic Conference on Informatics*, pages 448–456, 2005.

[6] P. Donmez, J. G. Carbonell, and P. N. Bennett. Dual strategy active learning. In *ECML*, pages 116–127, 2007.

[7] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, pages 97–112, 2002.

[8] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS 14*, pages 681–687. 2001.

[9] A. Esuli and F. Sebastiani. Active learning strategies for multi-label text classification. In *ECIR*, pages 102–113. 2009.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[11] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[12] S.-J. Huang, W. Gao, and Z.-H. Zhou. Fast multi-instance multi-label learning. In *CORR abs/1310.2049*, 2013.

[13] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. In *NIPS 23*, pages 892–900. 2010.

[14] C.-W. Hung and H.-T. Lin. Multi-label active learning with auxiliary learner. In *ACML*, pages 315–330, 2011.

[15] B. Klimt and Y. Yang. Introducing the enron corpus. In *Proceedinds of the 1st Conference on Email and Anti-Spam*, 2004.

[16] X. Li and Y. Guo. Active learning with multi-label svm classification. In *IJCAI*, 2013.

[17] X. Li, L. Wang, and E. Sung. Multilabel svm active learning for image classification. In *ICIP*, pages 2207–2210, 2004.

[18] H. T. Nguyen and A. W. M. Smeulders. Active learning using pre-clustering. In *ICML*, pages 623–630, 2004.

[19] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, and W. Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104, 2007.

[20] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. In *CVPR*, pages 1–8, 2008.

[21] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[22] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[23] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[24] M. Singh, A. Brew, D. Greene, and P. Cunningham. Score normalization and aggregation for active learning in multi-label classification. Technical report, University College Dublin, 2010.

[25] M. Singh, E. Curran, and P. Cunningham. Active learning for multi-label image annotation. In *Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science*, 2008.

[26] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *ICML*, pages 999–1006, 2000.

[27] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-label classification of music into emotions. In *Proceedings of the 9th International Conference of Music Information Retrieval*, page 325, 2008.

[28] N. Usunier, D. Buffoni, and P. Gallinari. Ranking with ordered weighted pairwise classification. In *ICML*, pages 1057–1064, 2009.

[29] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, pages 2764–2770, 2011.

[30] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *KDD*, pages 917–926, 2009.

[31] M.-L. Zhang and Z.-H. Zhou. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[32] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *TKDE*, in press.

[33] X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma. Multi-view multi-label active learning for image classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 258–261, 2009.