Multi-Label Active Learning: Query Type Matters

Sheng-Jun Huang[1,3] and **Songcan Chen**[1,3] and **Zhi-Hua Zhou**[2,3*]
[1]College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics
[2]National Key Laboratory for Novel Software Technology, Nanjing University
[3]Collaborative Innovation Center of Novel Software Technology and Industrialization
{huangsj, s.chen}@nuaa.edu.cn zhouzh@lamda.nju.edu.cn

Abstract

Active learning reduces the labeling cost by selectively querying the most valuable information from the annotator. It is essentially important for multi-label learning, where the labeling cost is rather high because each object may be associated with multiple labels. Existing multi-label active learning (MLAL) research mainly focuses on the task of selecting instances to be queried. In this paper, we disclose for the first time that the query type, which decides what information to query for the selected instance, is more important. Based on this observation, we propose a novel MLAL framework to query the relevance ordering of label pairs, which gets richer information from each query and requires less expertise of the annotator. By incorporating a simple selection strategy and a label ranking model into our framework, the proposed approach can reduce the labeling effort of annotators significantly. Experiments on 20 benchmark datasets and a manually labeled real data validate that our approach not only achieves superior performance on classification, but also provides accurate ranking for relevant labels.

1 Introduction

In many real world applications, there are plentiful unlabeled data but limited labeled data, and the acquisition of class labels is usually costly and difficult. By actively and iteratively selecting the most valuable data to query their supervised information, active learning tries to train an effective model with least labeling cost [Settles, 2009].

Multi-label learning deals with objects that are simultaneously associated with multiple labels [Zhang and Zhou, 2014], and has been successfully applied to various tasks, e.g., image classification [Bi and Kwok, 2013; Boutell *et al.*, 2004], text categorization [Schapire and Singer, 2000; Mencia and Furnkranz, 2008] and gene function prediction [Elisseeff and Weston, 2002]. To label a multi-label object, the human annotator needs to identify its relevance to every possible label, which leads to even higher cost than in single-label learning. Thus active learning for multi-label tasks has attracted more and more research interests in recent years.

Under the traditional single-label setting, active learning methods select the most valuable instances and then query their labels from the annotator (oracle). The key task is to design the criterion for instance selection [Settles, 2009]. Most research on multi-label active learning follows this principle, and focuses on selection criterion design [Esuli and Sebastiani, 2009; Singh *et al.*, 2010; Wang *et al.*, 2012]. For example, Vasisht et. al. [2014] proposed a mutual information based criterion for sparse Bayesian multi-label active learning; Brinker [2006] proposed to select instances according to the reduction of the version space volume; In both [Yang *et al.*, 2009] and [Hung and Lin, 2011], instances with maximum loss reduction are selected for query; while in [Singh *et al.*, 2008], the average uncertainty over all labels is used as the selection criterion. Besides, there are some works trying to combine multiple criteria for better selection [Li and Guo, 2013; Tang *et al.*, 2012; Huang *et al.*, 2014].

However, we observe that the query type, which decides what kind of supervised information to query for the selected instance, matters more than the selection criterion to the performance of multi-label active learning (MLAL). In fact, the expansion of label space in multi-label learning offers more potential options for the design of query types in MLAL. Given a selected instance, most existing methods query all the labels for the instance [Li *et al.*, 2004; Brinker, 2006; Chakraborty *et al.*, 2011; Hung and Lin, 2011; Li and Guo, 2013]. Recently, there are a few works trying to query the relevance of a instance-label pair, i.e., ask the oracle whether a specific label is relevant to the selected instance at each iteration [Qi *et al.*, 2008; Huang and Zhou, 2013; Huang *et al.*, 2014]. These simple query types suffers from the following shortcomings. First, querying all labels of an instance may lead to information redundancy and wasting of annotators' effort. It is well known that different labels are usually correlated in multi-label learning [Zhang and Zhou, 2014]. Thus only a part of labels need to be queried, while the rest can be inferred by exploiting the correlation among labels. Besides, in some real problems with a large number of candidate labels, annotators may hardly identify all relevant labels at a time. Querying a label-instance pair avoids information redundancy, but usually ignores the interaction be-

*This research was supported by the NSFC (61333014), 973 Program (2014CB340501) and China Postdoctoral Science Foundation.

tween labels, and gets limited supervision from each query. Moreover, both these two query types require the annotator to precisely decide the relevance of labels to instances, which may lead to high cost in complicated tasks. For example, in medical image analysis, only experts with rich experiences can accurately identify the disease of a patient based on the medical image. In contrast, if the question is to decide which of two given diseases is more likely suffered by the patient, then even medical students with basic knowledge may easily give the answer. So different query types may bring different information and cause different cost.

In this paper, we propose a multi-label active learning framework with a novel query type. Under this framework, we iteratively select one instance along with a pair of labels, and then query their relevance ordering, i.e., ask the oracle which of the two labels is more relevant to the instance. The proposed query type, on one hand reduces the requirements of annotators' expertise, and on the other hand, provides more useful information to improve the classification model. After each query, we employ a label ranking model with threshold learning for multi-label classification. The queried information is fully utilized to improve the classification model. With the proposed framework, we can incorporate with different selection criteria to implement different multi-label active learning algorithms. In this paper, we propose to select the least queried instance and most valuable label pairs for query. Our empirical study on 21 datasets demonstrates the advantage of the proposed query type as well as the efficacy of the implemented algorithm.

Our main contributions are summarized as follows.

- For the first time, we disclose that query type matters more to the performance of multi-label active learning than selection criterion.
- We propose a novel query type to ask for the relevance ordering of two labels on a specific instance, and further incorporate it with a new selection strategy to implement an effective multi-label active learning algorithm.
- In addition to superiority on classification performance, the proposed algorithm provides accurate ranking of relevant labels for unseen instances.

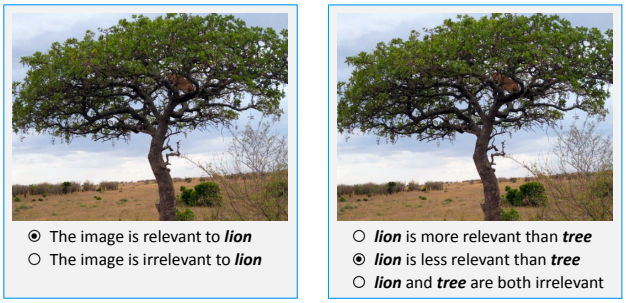
The rest of this paper is organized as follows. In Section 2, the proposed approach is introduced. Section 3 presents the experiments, followed by the conclusion in Section 4.

2 The Approach

In this section, we first present a multi-label active learning framework with the proposed query type, and then introduce the label ranking model for multi-label classification. After that, a matching criterion is proposed to select the labels and instances for query.

2.1 The framework

At each iteration, we query the relevance ordering of two labels to the selected instance. Denoting by $\mathbf{x}$, y_1 and y_2 the selected instance and two labels, respectively, the annotator needs to decide which of y_1 and y_2 is more relevant to $\mathbf{x}$. We offer three options: a) y_1 is more relevant than y_2; b) y_1 is less relevant than y_2; and c) both y_1 and y_2 are irrelevant to $\mathbf{x}$. From the answer a), we can know that y_1 is a relevant label of $\mathbf{x}$, while y_2 can be either an irrelevant label or a less relevant label. Similarly, we cannot directly know whether y_1 is a relevant label or not from the answer b). While for the answer c), both of the two labels are explicitly labeled as irrelevant. We will discuss how to utilize such supervised information for model training in the next subsection. Note here we assume that annotators always give a relative order for two relevant labels, and thus do not provide the option for the case that two labels are equally relevant. One can design more options to obtain more detailed information.



(a) an existing query type (b) the proposed query type

Figure 1: The interface of two query types for image annotation. (a) querying whether a label is relevant to the image; (b) querying which of two labels is more relevant to the image.

Figure 1 shows a comparison of two query types: (a) for the existing query type, i.e., querying the relevance of an instance-label pair, and (b) for the proposed type, i.e., querying the relevance ordering of two labels on an instance. To examine the efficiency of these two interfaces, we conduct a user study for the image annotation task. Each of three users is asked to answer 1000 queries via the two interfaces, respectively. Before the study, all the users are well trained to get familiar with the interface. Images and labels are randomly drawn from the *image* dataset [Zhang and Zhou, 2007]. Two of the three annotators finish the 1000 queries more quickly with the interface (b), and the average response time for each query with the two interfaces are very close, 1.38 seconds for the interface (a) and 1.43 seconds for the interface (b).

The proposed query type has two significant advantages. First, it does not require the annotator to provide exact labels for the selected instance. In some complicated applications such as medical image analysis, only experts with rich experience can give exact labels, while normal annotators can easily judge the relative order of the relevance of two labels. Even in common tasks such as image classification, the proposed query type still show its superiority. For example, for the image in Figure 1, people may overlook the lion in the tree, and mistakenly decide that *lion* is not a relevant label of the image. Instead, with the interface of our proposed query type, people can confidently identify that the image is more relevant to *tree* than *lion*. Another advantage of the proposed query type is that the supervised information obtained from each query explicitly contains the correlation between different labels, which has been shown to be crucial for multi-label

learning [Zhang and Zhou, 2014].

A framework of multi-label active learning with the proposed query type, termed as AURO (Active qUery on Relevance Ordering), can be summarized as follows. First, we select a triplet consisting of one instance $\mathbf{x}$ and two labels y_1 and y_2. Then, the relative order of y_1 and y_2 based on their relevance to the instance $\mathbf{x}$ is queried. After that, the queried information is utilized to update the classification model. This process is repeated until enough information is queried or the query cost reaches a specific threshold. In the following subsections, we will introduce the updating rules for the classification model and the selection strategy for the instances and labels to be queried.

2.2 Classification model

The algorithm recently proposed in [Huang and Zhou, 2013] is employed for multi-label classification in our AURO framework. In this section, we will first introduce this algorithm in detail. The basic idea is to rank relevant labels before all irrelevant ones by minimizing an approximated rank loss [Weston *et al.*, 2011]. Meanwhile, a dummy label is trained to separate relevant and irrelevant labels from the ranked label list. We denote by Y_i and $\bar{Y}_i$ the sets of relevant and irrelevant labels of the instance $\mathbf{x}_i$, respectively. The dummy label is denoted by y_0. A two-level model on label y is defined as

$$f_y(\mathbf{x}) = \mathbf{w}_y^\top W_0 \mathbf{x},$$

where W_0 maps the original feature vectors to a shared subspace, and then $\mathbf{w}_y$ builds a linear classifier on the subspace.

For an instance $\mathbf{x}$ and one of its relevant labels y, the ranking error can be defined as:

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

where

$$R(\mathbf{x}, y) = \sum_{\bar{y} \in \bar{Y}} I[f_{\bar{y}}(\mathbf{x}) > f_y(\mathbf{x})] \tag{2}$$

counts the number of irrelevant labels that are ranked before y. Obviously, the ranking error ϵ would be larger if y is lower ranked. Our target is to rank all labels correctly by minimizing the ranking error on all instances. Directly optimizing the non-convex error in Eq. 1 is a NP-hard problem. We thus introduce the hinge loss as a convex surrogate loss. By further decomposing the loss into all irrelevant labels, we can have Eq. 3 to define the loss induced by a relevant label y and an irrelevant label $\bar{y}$ on $\mathbf{x}$.

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

where $|q|_+ = \max(q, 0)$. Then the target is to minimize the ranking error on the training set:

$$\sum_{i=1}^{n} \sum_{y \in Y_i} \sum_{\bar{y} \in \bar{Y}_i} \mathcal{L}(\mathbf{x}_i, y, \bar{y}). \tag{4}$$

The algorithm iteratively samples data and employs stochastic gradient descent (SGD) to minimize the ranking error. At the t-th iteration of SGD, assuming the sampled triplet is $(\mathbf{x}, y, \bar{y})$, the model parameters can be updated according to:

$$W_0^{t+1} = W_0^t - \gamma_t \sum_{i=1}^{R(\mathbf{x},y)} \frac{1}{i} (\mathbf{w}_{\bar{y}}^t \mathbf{x}^\top - \mathbf{w}_y^t \mathbf{x}^\top) \tag{5}$$

$$\mathbf{w}_y^{t+1} = \mathbf{w}_y^t + \gamma_t \sum_{i=1}^{R(\mathbf{x},y)} \frac{1}{i} W_0^t \mathbf{x} \tag{6}$$

$$\mathbf{w}_{\bar{y}}^{t+1} = \mathbf{w}_{\bar{y}}^t - \gamma_t \sum_{i=1}^{R(\mathbf{x},y)} \frac{1}{i} W_0^t \mathbf{x} \tag{7}$$

where γ_t is the step size.

Note that the above introduced algorithm is designed for data with exact labels, and should be adapted to fully utilize the relevance ordering information queried in our AURO framework. Assume that in the current iteration of active learning, the relevance ordering of y_1 and y_2 to $\mathbf{x}$ is queried from the annotator. The sets of relevant and irrelevant labels of $\mathbf{x}$ known from previous queries are denoted as Y^+ and Y^-, respectively. Then with different answers from the annotator, we have different objective functions accordingly. If the answer is that y_1 is more relevant to $\mathbf{x}$ than y_2, then y_1 is added into Y^+ as a relevant label of $\mathbf{x}$, and the model should be trained with the following objectives. First, y_1 should be ranked before y_2. Second, $\mathbf{y}_1$ should be ranked before all irrelevant labels in Y^- as well as the dummy label y_0. The objective function can be written as:

$$\min \mathcal{L}(\mathbf{x}, y_1, y_2) + \sum_{\bar{y} \in Y^- \cup \{y_0\}} \mathcal{L}(\mathbf{x}, y_1, \bar{y}). \tag{8}$$

If the answer is that y_1 is less relevant to $\mathbf{x}$ than y_2, then y_2 is added into Y^+. We can have a similar objective function just by switching y_1 and y_2 in Eq. 8.

At last, if the answer is that both y_1 and y_2 are irrelevant to $\mathbf{x}$, then y_1 and y_2 are added into Y^-. We expect the model to rank both y_1 and y_2 behind the dummy label y_0 as well as all relevant labels in Y^+, leading to the following objective function:

$$\min \sum_{y \in Y^+ \cup \{y_0\}} \sum_{\bar{y} \in \{y_1, y_2\}} \mathcal{L}(\mathbf{x}, y, \bar{y}). \tag{9}$$

The above optimization problems can be efficiently solved via stochastic gradient descent.

2.3 Selection strategy

In this subsection, we show how the instance and labels are selected for query, respectively. First, the instance is selected based on uncertainty, which is a commonly used criterion. In this paper, we simply measure the uncertainty of an instance with the number of queries performed on it. The less queries have been performed, the more uncertain the instance is. Thus at each iteration of active learning, the least queried instance is selected. In case there are multiple instances with the same number of queries, we randomly pick one of them.

Then for the selected instance $\mathbf{x}$, we need further select two labels, denoted by y_1 and y_2, to query their relevance ordering. The two labels are expected to have the following properties. First, the current model should be less confident on the

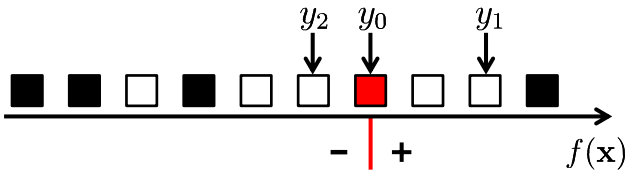


Figure 2: The selection strategy of labels y_1 and y_2. Labels are ranked based on the prediction values on instance $\mathbf{x}$. The black and white boxes indicate queried and not queried labels, respectively, while the red box represents the dummy label y_0, which separates relevant and irrelevant labels. y_1 and y_2 are the two labels selected according to Eq. 10 and Eq. 11

relevance ordering between y_1 and y_2. Otherwise querying some information that is already known by the model is useless for further improving it. Second, there should be a significant difference between the relevance of y_1 and y_2 to $\mathbf{x}$. Otherwise the queried information will induce only a slight change to the classification model. This suggests that the prediction on the two labels, i.e., $f_{y_1}(\mathbf{x})$ and $f_{y_2}(\mathbf{x})$, should not be too close. At last, relevant labels are preferred. Multi-label learning usually suffers from class-imbalance problem because each instance is relevant to only a small subset of all labels. Thus information on relevant labels is more valuable for improving the model, as suggested in [Huang and Zhou, 2013]. Taking all these factors into account, without loss of generality, y_1 and y_2 can be selected as:

$$y_1 = \underset{y \in U(\mathbf{x})}{\operatorname{argmax}} f_y(\mathbf{x}), \tag{10}$$

$$y_2 = \underset{y \in U(\mathbf{x})}{\operatorname{argmin}} |f_{y_0}(\mathbf{x}) - f_y(\mathbf{x})|, \tag{11}$$

where $U(\mathbf{x})$ denotes the set of labels not queried yet for $\mathbf{x}$. Note that Eq. 11 was firstly introduced in [Huang and Zhou, 2013]. The selection strategy is summarized in Figure 2. Note that the dummy label y_0 represents the boundary of the current model for separating relevant and irrelevant labels. y_2 is ranked after y_0, and thus is predicted as an irrelevant label. Based on the current model, y_1 is the most relevant label, and is expected to have a significant difference in relevance with the irrelevant label y_2. On the other hand, y_2 is the irrelevant label closest to the dummy label (the decision boundary), and thus its prediction is less confident.

It is worth noticing that the main contribution of this work is proposing AURO framework with the new query type. Within this framework, one can design various selection criteria to implement different algorithms. In this paper, we only present a simple strategy to select instances and labels. It is shown in the experiments that even with this simple strategy, our algorithm can achieve excellent performance owing to the superiority of the proposed query type.

3 Experiments

3.1 Settings

To evaluate the proposed approach, we compare the following six multi-label active learning algorithms in our experiments:

Table 1: Statistics on data sets. LC (label cardinality) is the average number of relevant labels per instance.

Data	# instance	# label	# feature	LC
Corel5K	5,000	374	499	3.52
Emotions	593	6	72	1.87
Enron	1,702	53	1,001	3.38
Genebase	662	27	1,185	1.25
Image	2,000	5	294	1.24
Medical	978	45	1,449	1.25
Reuters	2,000	7	243	1.15
Scene	2,407	6	294	1.07
Yeast	2,417	14	103	4.24
Arts	5,000	26	462	1.64
Business	5,000	30	438	1.59
Computers	5,000	33	681	1.51
Education	5,000	33	550	1.46
Entertainment	5,000	21	640	1.42
Health	5,000	32	612	1.66
Recreation	5,000	22	606	1.42
Reference	5,000	33	793	1.17
Science	5,000	40	743	1.45
Social	5,000	39	1,047	1.28
Society	5,000	27	636	1.69

- MMC: the algorithm proposed in [Yang *et al.*, 2009], which selects instances based on the loss reduction criterion.
- Adaptive: the adaptive method proposed in [Li and Guo, 2013], which selects instances based on both the max-margin prediction uncertainty and the label cardinality inconsistency.
- AUDI: the algorithm proposed in [Huang and Zhou, 2013], which selects instance-label pairs based on uncertainty and diversity.
- QUIRE: the algorithm proposed in [Huang *et al.*, 2014], which selects instance-label pairs by simultaneously considering informativeness and representativeness.
- AURO-r: a baseline implementation of our AURO framework, which queries the relevance ordering of randomly selected labels and instances.
- AURO: the algorithm developed in this paper, which incorporates the proposed query type and selection strategy.

These approaches employ different query types. At each iteration, MMC and Adaptive query all labels for one instance; AUDI and QUIRE query the relevance of an instance-label pair; while AURO-r and AURO query the relevance ordering of two labels. As stated in the last section, querying an instance-label pair and querying the relevance ordering of two labels cost almost the same time according to our user study. Although the proposed query type requires less expertise of annotators, we simply use the number of queries as the measurement of query cost for convenience. It is worth noticing that querying all labels for one instance is the equivalent of

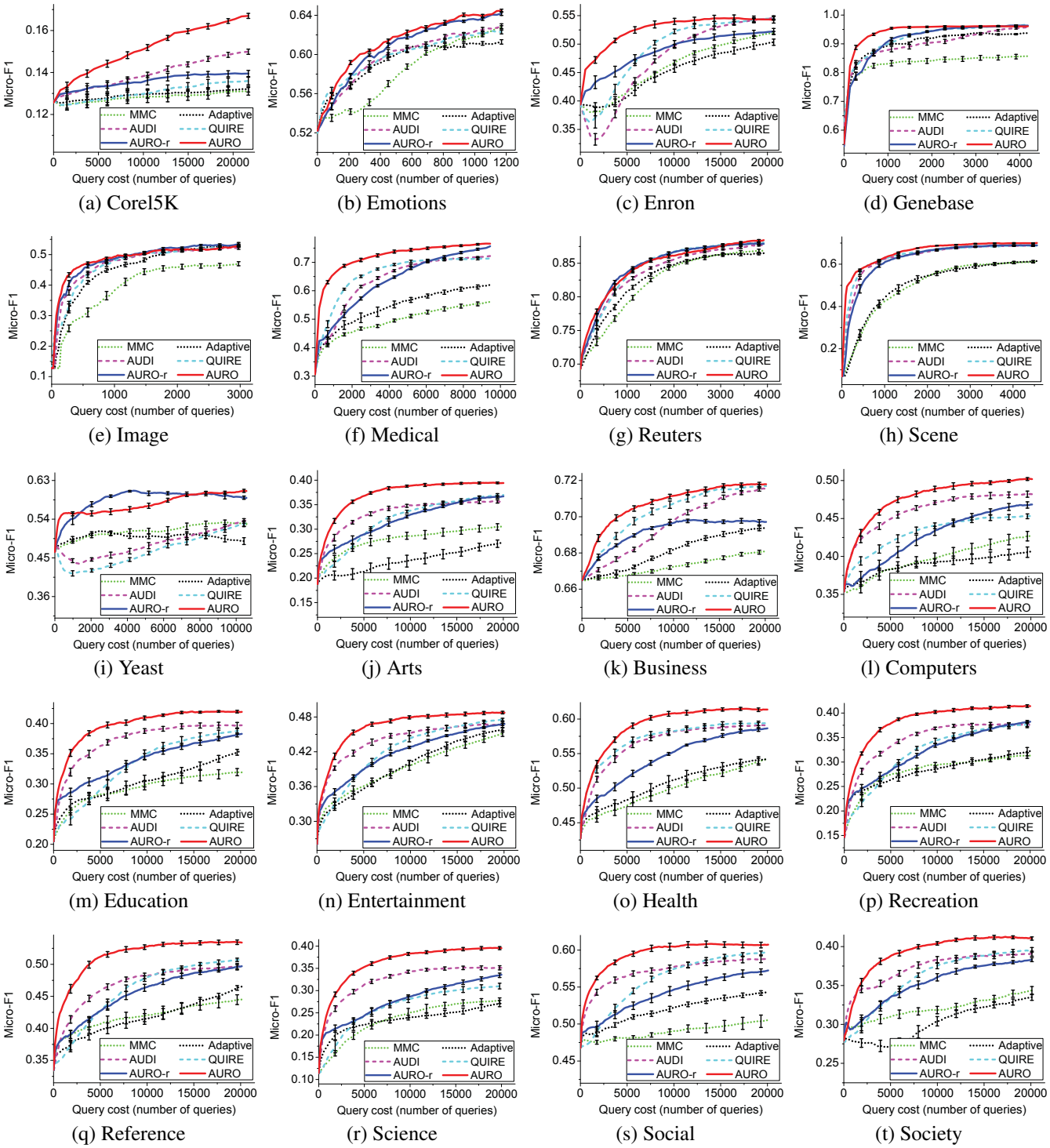


Figure 3: The average and standard deviation results in terms of *Micro-F1* on 20 datasets.

querying m instance-label pairs, because it requires the annotator to identify the relevance of the instance to every possible labels. Here m denotes the number of all candidate labels.

We use the multi-label learning algorithm proposed in the AUDI work [Huang and Zhou, 2013] as the classification model for all compared approaches. For each experiment, we randomly divide the dataset into three parts: the test set with 50% examples, the initial labeled set with 5% examples and the unlabeled pool with the rest instances. Parameters are selected via leave-one-out cross validation on the initial

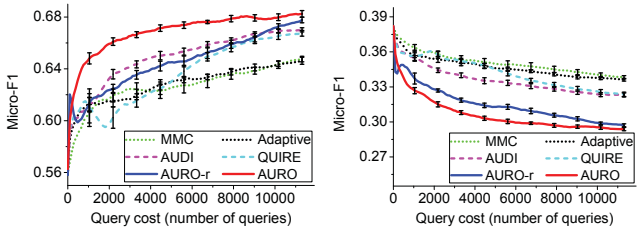
set. After every $5 \times m$ queries, we evaluate the performance of the classification model on the test set. The querying process is stopped if all data are labeled or the number of queries reaches 20,000. We repeat the each experiment for 10 times and report the average results.

3.2 Results on benchmark datasets

We first perform the experiments on 20 benchmark datasets. The statistical information of these datasets are summarized in Table 1. The data size varies from 593 to 5,000, while the number of labels varies from 5 to 499. Note that we are using the original corel5K dataset, and thus the results may be inconsistent with those reported in other literatures, where subset of corel5K was used. We evaluate the performance with micro-F1, which is a commonly used performance measure in multi-label learning. The comparison results are shown in Figure 3. We use different line styles to represent different query types in the figures: dotted line for querying all labels of one instance, dashed line for querying instance-label pairs, and solid line for querying relevance ordering of label pairs.

As observed from Figure 3, our approach AURO achieves the best performance on all datasets. In general, AUDI and QUIRE are superior to MMC and Adaptive. This indicates that querying instance-label pairs is more effective than querying all labels for one instance. While compared to our approach, both of these two query types are outperformed by the proposed query type. It is worthy to note that AURO-r, which combines our query type with random selection, surprisingly achieves comparable performance to state-of-the-art methods, or even outperforms MMC and Adaptive in most cases. These results demonstrate that an effective query type with simple or even random selection strategy can achieve excellent performance, and validate our conclusion that the query type matters more to the performance of multi-label active learning than selection criterion. For the two methods querying all labels of an instance, Adaptive is more effective than MMC, probably because it adaptively uses multiple criteria to select instances. The results of AUDI and QUIRE are comparable, while AUDI tends to be superior on more datasets. At last, when comparing AURO with AURO-r, it can be found that they achieve comparable results on *image* and *reuters*; while on the other datasets, AURO outperforms AURO-r in most cases. In general, the performances of AURO-r and AURO are close on datasets with fewer labels, possibly because the selection criterion in Eq. 11 is more distinguishable when there are more labels.

To examine the significance of the results, we compare AURO with each competing method after each query, and then count the times of our win/tie/loss based on t-test at 95% confidence level. Due to space limitation, here we report only the average counts (in percentage) on all the data sets. The win/tie/loss counts (%) of AURO are $98/2/0$ versus MMC, $97/3/0$ versus Adaptive, $90/10/0$ versus AUDI, $87/13/0$ versus QUIRE and $83/14/3$ versus AURO-r. These results show that the proposed AURO approach has significant advantage over the compared methods.



(a) Micro F1 (b) ProLoss

Figure 4: Comparison results on *MSRA* in terms of micro-F1 and ProLoss.

3.3 Results on MSRA

Experiments on benchmark datasets have validated the superiority of the proposed approach on classification performance. In this subsection, we further study the effectiveness of our approach on ranking relevant labels. As proposed in Xu and Zhou's work [2013], in addition to differentiating relevant labels from irrelevant ones, ranking the relevant labels is desired in many tasks. MSRA is a multi-label dataset for image classification, and consists of 1868 images with 19 candidate labels. The ordering of relevant labels has been manually provided for each instance. Thus we can evaluate the ranking quality of different algorithms by comparing their predicted ranking with the groundtruth ranking. In addition to micro-F1, ProLoss [Xu *et al.*, 2013], which measures both the prediction accuracy and the ranking quality of relevant labels, is employed as the performance measure. Note that for ProLoss, a smaller value indicates a better performance.

The comparison results are presented in Figure 4. As we can see, our approach AURO achieves the best performance on both micro-F1 and ProLoss, while the advantage on ProLoss is rather significant. Especially, AURO-r outperforms all the other methods on ProLoss. These results validated that our approach not only well separates relevant and irrelevant labels, but also provides accurate ranking of relevant labels.

4 Conclusion

Existing research on multi-label active learning research mainly focuses on designing criteria for selecting instances to be queried. In this paper, for the first time, we disclose that the query type matters more than the selection criterion to the performance of MLAL. We propose a novel framework to query the relevance ordering of two labels on a specific instance, and further incorporate it with a new selection strategy to implement an effective MLAL algorithm. The proposed approach, on one hand reduces the labeling efforts of annotators and achieves superior classification performance to state-of-the-art methods; and on the other hand, can rank the relevant labels accurately for unseen instances. Extensive study on more than 20 datasets validated the effectiveness of our approach. In the future, we plan to design other selection strategies to further improve our approach. In addition, novel query types, e.g., querying relative relevance of instances pairs, will be studied.

References

[Bi and Kwok, 2013] W. Bi and J. Kwok. Efficient multi-label classification with many labels. In *Proceedings of the 30th International Conference on Machine Learning*, pages 405–413, Atlanta, GA, 2013.

[Boutell *et al.*, 2004] M.R. Boutell, J. Luo, X. Shen, and C.M. 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 *Proceedings of the 29th Annual Conference of the Gesellschaft for classification*, pages 206–213, Magdeburg, Germany, 2006.

[Chakraborty *et al.*, 2011] 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, Scottsdale, AZ, 2011.

[Elisseeff and Weston, 2002] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14*, pages 681–687. 2002.

[Esuli and Sebastiani, 2009] A. Esuli and F. Sebastiani. Active learning strategies for multi-label text classification. In *Proceedings of the 31th European Conference on Information Retrieval Research*, pages 102–113, Toulouse, France, 2009.

[Huang and Zhou, 2013] S.-J. Huang and Z.-H. Zhou. Active query driven by uncertainty and diversity for incremental multi-label learning. In *Proceedings of the 13th IEEE International Conference on Data Mining*, pages 1079–1084, Dallas, TX, 2013.

[Huang *et al.*, 2014] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (10):1936–1949, 2014.

[Hung and Lin, 2011] C.-W. Hung and H.-T. Lin. Multi-label active learning with auxiliary learner. In *Proceedings of the 3rd Asian Conference on Machine Learning*, pages 315–330, Taoyuan, Taiwan, 2011.

[Li and Guo, 2013] X. Li and Y. Guo. Active learning with multi-label svm classification. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1479–1485, Beijing, China, 2013.

[Li *et al.*, 2004] X. Li, L. Wang, and E. Sung. Multilabel svm active learning for image classification. In *Proceedings of the 11th International Conference on Image Processing*, pages 2207–2210, Singapore, 2004.

[Mencia and Furnkranz, 2008] E. L. Mencia and J. Furnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Proceedings of the 19th European Conference on Machine Learning*, pages 50–65, Antwerp, Belgium, 2008.

[Qi *et al.*, 2008] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, 2008.

[Schapire and Singer, 2000] R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168, 2000.

[Settles, 2009] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2009.

[Singh *et al.*, 2008] 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*, pages 173–182, Dublin, Ireland, 2008.

[Singh *et al.*, 2010] 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.

[Tang *et al.*, 2012] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua. Semantic-gap-oriented active learning for multilabel image annotation. *IEEE Transactions on Image Processing*, 21(4):2354–2360, 2012.

[Vasisht *et al.*, 2014] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 472–481, New York, NY, 2014.

[Wang *et al.*, 2012] P. Wang, P. Zhang, and L. Guo. Mining multi-label data streams using ensemble-based active learning. In *Proceedings of the 12th SIAM International Conference on Data Mining*, pages 1131–1140, Anaheim, CA, 2012.

[Weston *et al.*, 2011] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 2764–2770, Barcelona, Spain, 2011.

[Xu *et al.*, 2013] M. Xu, Y.-F. Li, and Z.-H. Zhou. Multi-label learning with pro loss. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 998–1004, Bellevue, WA, 2013.

[Yang *et al.*, 2009] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 917–926, Paris, France, 2009.

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

[Zhang and Zhou, 2014] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819, 2014.