Cost Effective Multi-label Active Learning via Querying Subexamples

Xia Chen[1,3], Guoxian Yu[1,*], Carlotta Domeniconi[2], Jun Wang[1], Zhao Li[3], Zili Zhang[1,4]
[1]College of Computer and Information Sciences, Southwest University, China
[2]Department of Computer Science, George Mason University, USA
[3]Alibaba Group, Hangzhou, China
[4]School of Information Technology, Deakin University, Australia
Email: [1]{xchen, gxyu, kingjun, zhangzl}@swu.edu.cn; [2]carlotta@cs.gmu.edu; [3]lizhao.lz@alibaba-inc.com

***Abstract*—Multi-label active learning addresses the scarce labeled example problem by querying the most valuable unlabeled examples, or example-label pairs, to achieve a better performance with limited query cost. Current multi-label active learning methods require the scrutiny of the whole example in order to obtain its annotation. In contrast, one can find positive evidence with respect to a label by examining specific patterns (i.e., subexample), rather than the whole example, thus making the annotation process more efficient. Based on this observation, we propose a novel two-stage cost effective multi-label active learning framework, called CMAL. In the first stage, a novel example-label pair selection strategy is introduced. Our strategy leverages label correlation and label space sparsity of multi-label examples to select the most uncertain example-label pairs. Specifically, the unknown relevant label of an example can be inferred from the correlated labels that are already assigned to the example, thus reducing the uncertainty of the unknown label. In addition, the larger the number of relevant examples of a particular label, the smaller the uncertainty of the label is. In the second stage, CMAL queries the most plausible positive subexample-label pairs of the selected example-label pairs. Comprehensive experiments on multi-label datasets collected from different domains demonstrate the effectiveness of our proposed approach on cost effective queries. We also show that leveraging label correlation and label sparsity contribute to saving costs.**

***Index Terms*—multi-label learning, active learning, cost effective, label correlations**

I. Introduction

Multi-label active learning, which is more challenging than single-label active learning, has attracted an enormous amount of research. Some multi-label active learning approaches follow the standard example-based selection criteria, and simultaneously query all labels of the selected examples [1], [2]. If the label space is large, querying all relevant labels of an example becomes expensive and time-consuming, since the annotator needs to traverse the whole label space. Other approaches attempt to select specific example-label pairs and then query their relevance [3], [4], [5]. It is recognized that an appropriate example-label pair selection strategy can save the labeling cost to a great extent.

Whether or not a particular label is relevant for an example depends on the characteristics of the example itself [6], [7]. For example, documents are typically organized in paragraphs. Current active learning approaches query a document with respect to a particular category or topic, requiring the annotator to go through the whole document (i.e., all paragraphs). However, the annotator could more easily annotate the document by browsing one (or selected) paragraph(s), instead of the whole document, once the paragraph more likely to be relevant to the topic has been selected. It's obvious that annotating a paragraph is more cost-saving than annotating the whole document. In other words, querying the *whole* example-label pair may result in *information redundancy* and in a waste of resources, if its subexamples (i.e., paragraphs) related to the label can be selected. Thus, we consider querying the most likely positive subexample-label pairs instead of the example-label pair. In addition, label correlation and sparsity of the label space are crucial to multi-label learning [8], [9]. By leveraging label correlation, the unknown relevant labels of an example can be inferred from the correlated labels that are already assigned to the example. Label space sparsity implies that a label is usually relevant to a small number of examples, or in other words, only a small number of labels is relevant to an example.

In light of the above observations, in this paper, we propose a novel multi-label active learning framework, called CMAL. CMAL makes use of label correlation and label space sparsity, and it iteratively queries the most likely positive subexample-label pairs, instead of the whole example-label pairs. We simply assume that the cost (e.g., effort) of querying an example-label pair is larger than that of querying its subexample-label pair. In each iteration, CMAL first measures how uncertain the example-label pairs are, by leveraging label correlation and label space sparsity. Then, CMAL queries the most uncertain example-label pair by querying its most likely positive subexample-label pairs, thus reducing the query cost. Experimental results on publicly available datasets show that CMAL achieves a performance comparable to other related methods at a smaller cost [5], [10], [11].

II. Related Work

Although traditional active learning has been widely studied in many application domains and gives good classification models at a small cost, in recent years, many researchers have

[*]Corresponding author: gxyu@swu.edu.cn (Guoxian Yu)

investigated effective active learning which can further reduce the annotation cost. Some methods [13], [14], [15] assume that multiple annotators are available to provide labels of different quality and cost, and then focus on designing active learning criteria to select example-labeler pairs. Unlike traditional active learning methods, which only focus on querying uncertain examples with low prediction confidence, some methods identify the majority of examples with highly confident predictions from the unlabeled set, and then automatically assign pseudo-labels to them without any human effort [16], [17]. Instead of querying the example or example-label pairs, Huang *et al.* [18] introduced a novel query type, where the relevance ordering of target label pairs with respect to a selected example is queried. This strategy has the advantage of obtaining richer information and requires less expertise from the annotator. In contrast, in this paper, we study cost effective active learning based on the *cost margin* between querying an example and its subexamples.

Considering the multiple subexamples of an example, our work is closely related to multi-instance learning, where each example is represented as a bag that contains multiple instances (or subexamples) [19]. In multi-instance learning, a bag is positive with respect to a particular label if at least one of its instances is positive, while the bag is negative if all its instances are negative for that label [19], [20]. Although active learning has been heavily studied for the single-instance scenario, only few methods have been proposed to address the multi-instance active learning problem, and our approach is different from all of them. Some methods assume that all bags are labeled and the learner is allowed to query the labels of instances from the *positive* bags to train an instance classifier [21] or a bag classifier [22]. In contrast, our method tries to query the labels of unlabeled bags (i.e., examples) for bag classification. The learning scenarios in [23], [24] also aim at querying the labels of unlabeled bags for bag classification. However, these two methods query the labels of unlabeled bags based on all its instances, while our proposed method queries bag labels by querying some instance-label pairs of unlabeled bags and has a significantly reduced query cost. In [25], only one query round is used, and the annotator is required to annotate a region (not an example), which is the group of the most valuable instances. In addition, the aforementioned multi-instance active learning methods study binary classification, which is a special case of the multi-label classification problem studied in this paper. Although [10] and [11] investigated multi-label active learning under the case of multi-instances, they focus on querying all possible labels of unlabeled bags or the relevance of bag-label pairs based on all its instances, and have a cost that is higher than our solution.

III. Cost Effective Multi-label Active Learning

A. Problem Formulation

Let $\mathcal{L} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{l}$ be a small training set with l labeled examples, where $\mathbf{X}_i \in \mathbb{R}^d$ is the feature vector for the i-th example in the d-dimensional feature space, and $\mathbf{Y}_i \in \{-1, +1\}^q$ encodes the ground-truth labels of the i-th example in the q-dimensional label space. The value $\mathbf{Y}_{ic} = +1$ means that the example $\mathbf{X}_i$ is relevant to the c-th label, otherwise $\mathbf{Y}_{ic} = -1$. Let $\mathcal{U} = \{\mathbf{X}_j\}_{j=l+1}^{l+u}$ be a large pool of u unlabeled examples, where typically $l << u$. $\mathcal{B}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{in_i}\}$ is the set of subexamples (or instances) of $\mathbf{X}_i$, where $n_i = |\mathcal{B}_i|$ is the number of subexamples of $\mathbf{X}_i$. Each subexample $\mathbf{x}_{ik} \in \mathbb{R}^{d'}$ of an example $\mathbf{X}_i \in \mathcal{L}$ is associated with labels $\mathbf{y}_{ik} \in \{-1, +1\}^q$. As before, the value $\mathbf{y}_{ikc} = +1$ means that the subexample $\mathbf{x}_{ik}$ is relevant to the c-th label, otherwise $\mathbf{y}_{ikc} = -1$. $\mathcal{L}_{sub}$ represents the set of training subexamples from $\mathcal{L}$, and $\mathcal{U}_{sub}$ represents the unlabeled subexamples from $\mathcal{U}$. As in standard multiple instance learning, an example (or bag) is positive if it has at least one positive subexamples (or instances), while an example is negative if all its subexamples are negative.

A simple and effective solution for multi-label classification is to transform a multi-label classification problem into q independent binary classification problems (one per label) via the "one-vs-all" scheme. In this work, we conduct multi-label classification under such a scheme by utilizing SVM and multi-instance SVM (mi-SVM) [26] for example classification and instance classification, respectively. We train q SVM classifiers, f_1, f_2,..., f_q, and q mi-SVM classifiers, g_1, g_2,..., g_q, for examples and subexamples, respectively. Each classifier corresponds to a particular label. Eventually, $\hat{\mathbf{Y}}_{ic} = sign(f_c(\mathbf{X}_i))$ and $\hat{\mathbf{y}}_{ikc} = sign(g_c(\mathbf{x}_{ik}))$ are the predictions of example $\mathbf{X}_i$ and subexample $\mathbf{x}_{ik}$ for the c-th label, respectively.

B. Example and Subexample Selection Strategy

In this section, we introduce how the example-label pairs and subexample-label pairs are selected for asking queries. CMAL adopts the widely used uncertainty criterion to select an example-label pair, where the uncertainty reflects the confidence of the current classifier in classifying the example. A larger uncertainty value implies a lower confidence, and it indicates a larger informativeness of the selected example-label pair towards improving the classifier [27]. For SVMs, the uncertainty on the prediction of an example-label pair can be measured by the distance between the example and the decision boundary; a smaller distance implies a larger uncertainty. For the example-level SVM, the decision boundary with respect to the c-th label is defined as $f_c(\mathbf{X}) = 0$, and the distance between $\mathbf{X}_i$ and that label is defined as $|f_c(\mathbf{X}_i)|$. We convert the distance to a measure of uncertainty as follows:

$$U(\mathbf{X}_i, c) = \frac{1}{|f_c(\mathbf{X}_i)|} \tag{1}$$

Label correlation plays an important role in multi-label learning [8], [9]. Under the multi-label active learning scenario, the information embedded in a unknown label of an example can be inferred from its correlated labels that have been queried as relevant for the same example. Thus, incorporating label correlation into multi-label active learning, can reduce the number of queries and save cost. In addition, multi-label datasets are generally affected by the label space sparsity problem, i.e., only a small fraction of examples are relevant for a given label. Given these observations, we propose a weighted uncertainty based example-label pair selection strategy, which leverages both label correlation and label space sparsity. Specifically, the pairwise label correlation matrix $\mathbf{W}$ is estimated as follows:

$$\mathbf{W}(c_1, c_2) = \frac{\sum_{i=1}^{l+u}[\mathbf{Y}_{ic_1} = +1, \mathbf{Y}_{ic_2} = +1]}{\sum_{i=1}^{l+u}[\mathbf{Y}_{ic_1} = +1 \ \mathit{or} \ \mathbf{Y}_{ic_1} = -1, \mathbf{Y}_{ic_2} = +1]} \quad (2)$$

where $[x] = 1$ if x is true, $[x] = 0$ otherwise. $\mathbf{W} \in \mathbb{R}^{q \times q}$ measures the empirical conditional probability that an example is relevant to label c_1, given that the example was already found to be relevant to label c_2. Note that the estimation of the conditional probability $\mathbf{W}(c_1, c_2)$ is based on the examples whose relevance towards c_1 and c_2 is already known. This label correlation is widely adopted in multi-label learning for its simplicity and intuitiveness [28], [29]. Other estimations of label correlation can also be adopted [8]. To mitigate the impact of limited labels, $\mathbf{W}$ is iteratively updated as labels are queried during active learning.

The weighted uncertainty function is defined as:

$$\tilde{U}(\mathbf{X}_i, c) = w_1(\mathbf{X}_i, c) * w_2(c) * U(\mathbf{X}_i, c) \quad (3)$$

where

$$w_1(\mathbf{X}_i, c) = 1 - \frac{\sum_{c_1 \in \mathcal{Y}^+(\mathbf{X}_i)} \mathbf{W}(c, c_1)}{q}, w_2(c) = 1 - \frac{n_c^+}{n} \quad (4)$$

$w_1(\mathbf{X}_i, c) \in [0, 1]$ denotes the weight of the example-label pair $(\mathbf{X}_i, c)$, and $w_2(c) \in [0, 1]$ is the weight of the c-th label. $\mathcal{Y}^+(\mathbf{X}_i)$ is the set of queried relevant labels of $\mathbf{X}_i$; $n = l + u$ is the total number of examples in the training set and in the unlabeled pool; and n_c^+ is the number of positive examples of label c. $w_1(\mathbf{X}_i, c)$ is driven by the fact that a large $\mathbf{W}(c, c_1)$ means that c is correlated with c_1, and therefore c is also possibly relevant to $\mathbf{X}_i$, given that $\mathbf{X}_i$ is already annotated with c_1. As a consequence, the larger the correlation between c and labels that have been deemed relevant to $\mathbf{X}_i$ is, the less uncertain $(\mathbf{X}_i, c)$ is. Furthermore, the larger the number of labels that have been queried for an example is, the less uncertain the example is. From the definition of $w_2(c)$ we can see that, the larger the number of positive examples of label c, the smaller the uncertainty associated to c is.

Based on the definition of weighted uncertainty given in Eq. (3), CMAL selects the most uncertain example-label pair $(\mathbf{X}_{i^*}, c^*)$ as follows:

$$(\mathbf{X}_{i^*}, c^*) = \arg\max_{\mathbf{X}_i \in \mathcal{U}, c \in \mathcal{Q}_l(\mathbf{X}_i)} \tilde{U}(\mathbf{X}_i, c) \quad (5)$$

where i^* and c^* are the selected example and label indeces, respectively; and $\mathcal{Q}_l(\mathbf{X}_i)$ is the set of not-queried labels for $\mathbf{X}_i$.

After this step, CMAL selects the most likely positive subexample-label pair of the selected example-label pair as follows:

$$(\mathbf{x}_{i^*k^*}, c^*) = \arg\max_{\mathbf{x}_{i^*k} \in \mathcal{Q}_{sub}(\mathcal{B}_{i^*})} g_{c^*}(\mathbf{x}_{i^*k}) \quad (6)$$

where k^* is the selected subexample index of example $\mathbf{X}_{i^*}$ and $\mathcal{Q}_{sub}(\mathcal{B}_{i^*})$ denotes the set of not-queried subexamples of the example-label pair $(\mathbf{X}_{i^*}, c^*)$. CMAL then queries the relevance of the selected subexample-label pair $(\mathbf{x}_{i^*k^*}, c^*)$.

C. Updating Strategy

In each iteration, after querying the subexample-label pair $(\mathbf{x}_{i^*k^*}, c^*)$, the response is either positive (relevant, +1) or negative (irrelevant, -1). If the feedback is *positive*, CMAL executes the following operations:

(i) $(\mathbf{X}_{i^*}, c^*)$ is annotated with +1, removed from $\mathcal{U}$ and added to the labeled example set $\mathcal{L}$.
(ii) $(\mathbf{x}_{i^*k^*}, c^*)$ is annotated with +1, removed from $\mathcal{U}_{sub}$ and added to the training example set $\mathcal{L}_{sub}$.
(iii) The current two classification models $f_{c^*}(\mathbf{X})$ and $g_{c^*}(\mathbf{x})$ are updated with the new $\mathcal{L}$ and $\mathcal{L}_{sub}$, respectively.

The example-label pair $(\mathbf{X}_{i^*}, c^*)$ is negative if all its subexample-label pairs have been queried and in each case the feedback was negative. Thus, for negative feedback, CMAL executes the following operations:

(i) If $(\mathbf{x}_{i^*k^*}, c^*)$ is the last being queried, and in each case the feedback was negative, then $(\mathbf{X}_{i^*}, c^*)$ is annotated with -1, removed from $\mathcal{U}$ and added to $\mathcal{L}$.
(ii) $(\mathbf{x}_{i^*k^*}, c^*)$ is annotated with -1, removed from $\mathcal{U}_{sub}$ and added to $\mathcal{L}_{sub}$.
(iii) $g_{c^*}(\mathbf{x})$ is updated based on the new $\mathcal{L}_{sub}$. If (i) is executed, then $f_{c^*}(\mathbf{X})$ is updated based on the new $\mathcal{L}$.

CMAL iteratively selects the most uncertain example-label pair and then queries the most likely positive subexample-label pair of the selected example-label pair. A positive example-label pair has at least one positive subexample-label pair. If the selected example-label pair is positive, it's expected that the learner will find the positive subexample-label pair early in the process, thus greatly reducing the query cost. If the selected example-label pair is negative, the learner can uncover the negative relationship by iteratively querying all its subexample-label pairs. Additional query strategies for subexample-label pairs of the selected negative example-label pair will be investigated in Section IV-D.

IV. Experiments

A. Experimental Setup

Datasets: To examine the effectiveness of the proposed method, we need the instance-level labels to simulate the oracle (annotator). Among the publicly available multi-instance multi-label datasets, four eligible datasets can be used for our experiments (summarized in Table I). They were previously used in multi-instance multi-label learning and active learning [30], [11]. Since the feature vectors of examples are unavailable for the four datasets, we followed the widely-adopted solution in [7] to generate them. In this paper, bags and instances are called examples and subexamples respectively.

Comparing Methods: We compare CMAL against the following methods:

(i) QUIRE [5] selects the most valuable example-label pair based on both informativeness and representativeness.

TABLE I
SUMMARY OF BENCHMARK DATASETS. AVGLABELS IS THE AVERAGE NUMBER OF LABELS PER BAG.

						Subexamples of each example		
Dataset	Example	Subexamples	Features	Labels	AvgLabels	Min	Avg	Max
Birds	548	10232	38	13	2.1	2	18.7	43
MSRC-v2	591	1758	48	23	2.5	1	3.0	17
Letter_carroll	166	717	16	26	3.9	1	4.3	12
Letter_frost	144	565	16	26	3.6	1	3.9	11

(ii) MidSelect [10] first transforms multi-instance multi-label examples into single-instance representations, and then applies the multi-label active learning strategy that selects the most uncertain examples to be queried.

(iii) MIML-AL [11] first selects the most valuable example-label pair based on diversity and uncertainty. To receive more specific supervision, it additionally requires the annotator to indicate the key subexample of this example-label pair once the pair is deemed positive.

(iv) CMAL-RM is a variant of CMAL; it randomly selects the example-label pair, and then selects the most likely positive subexample-label pair of the selected example-label pair for the query. This variant is used to study the effectiveness of selecting uncertain example-label pairs.

(v) CMAL-RR is another variant of CMAL; it first randomly selects the example-label pair and then randomly chooses subexample-label pairs of the selected example-label pair for the query.

To enable quantitative comparison, we follow the widely used assumption that querying all q labels for one example is equivalent to querying q example-label pairs [3], [4], [5]. Similarly, we assume that querying all subexamples of one example is equivalent to querying the whole example. This assumption is meaningful in crowdsourcing, where one divides a complex task into multiple simple micro tasks, and then distributes these tasks to crowdsourced workers. The total cost for addressing the micro tasks can be considered equal to the cost for addressing the overall complex task, which is hard or even infeasible to address as a whole. For a fair comparison, q one-vs-rest SVM (implemented with LIBSVM [31]) is used as the basic classification model to evaluate all the approaches. The parameters C (penalty of the error term) and γ (RBF kernel parameter) of SVMs used in our experiments are selected via 5-fold cross-validation on the initial set. For each experiment, we randomly divide the dataset into three parts: the test set with 50% of the examples, the initial labeled set with 5% of the examples, and the unlabeled pool with the rest of the examples. The subexamples of all the examples are also accordingly divided into three parts. We repeat the random data partition 10 times, and report the average results. After each query, we update the classification model on the extended labeled data and evaluate its performance on the holdout test set. The query process is stopped when all the examples in the unlabeled pool have been labeled.

Evaluation Metrics: We adopt the representative and widely-used multi-label learning evaluation metrics: *AveragePrecision* [8]. The larger the values of AveragePrecision, the better the performance is.

B. Comparison against State-of-the-art Methods

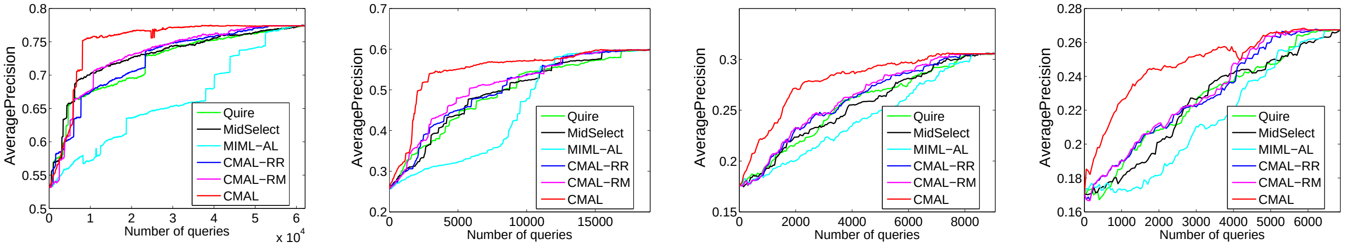
Figure 1 shows the *AveragePrecision* of all the methods on the four datasets, as a function of the number of queried subexample-label pairs (also equal to the total query cost).

From the figures, we can make the following observations. (i) CMAL significantly outperforms the other methods. This demonstrates the effectiveness of our proposed multi-label active learning strategy. (ii) CMAL-RM outperforms CMAL-RR in most cases, which demonstrates the effectiveness of selecting the most likely positive instance for saving query cost. In most cases, CMAL-RM and CMAL-RR achieve superior or comparable performance with respect to Quire, MidSelect, and MIML-AL. This observation further suggests that querying the relevance of instance-label pairs is more effective than querying the examples, or example-label pairs. (iii) MIML-AL often gives the lowest *AveragePrecision* values. The reason is that MIML-AL is a multi-instance multi-label active learning method, and it doesn't leverage feature information at the example level, while all the other methods do. (iv) To obtain a detailed supervision, MIML-AL asks the annotator to identify the key instance for the queried example-label pair which received positive feedback, but it mainly focuses on example-label pair queries, so it is also outperformed by CMAL. From these results, we can conclude that performing multi-label active learning by querying subexamples of the selected example-label pair can achieve a superior or comparable performance with less cost than other methods that query examples or example-label pairs.

C. Study on the Impact of Label Correlation and Label Space Sparsity

In this section, we conduct another experiment to study the contribution of label correlation and label space sparsity on multi-label active learning. To this end, we introduce three variants of CMAL: CMAL-nC, CMAL-nS and CMAL-nCS. CMAL-nC only employs label space sparsity to weight uncertainty, without considering label correlation. CMAL-nS only employs label correlation to weight uncertainty, without considering label space sparsity. CMAL-nCS directly uses the original uncertainty of example-label pairs without any weighting. Figure 2 gives the *AveragePrecision* of CMAL and the three variants on two datasets.

From the figure, we observe the following: (i) CMAL generally achieves a performance superior to CMAL-nS and CMAL-nCS. This shows that incorporating label space sparsity and label correlation helps the selection of more



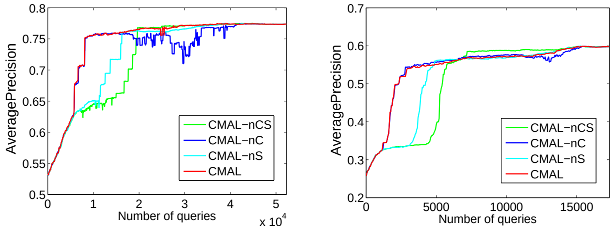
(a) AveragePrecision (b) AveragePrecision (c) AveragePrecision (d) AveragePrecision

Fig. 1. AveragePrecision vs. total number of queries of subexample-label pairs.

valuable example-label pairs. (ii) CMAL-nS often outperforms CMAL-nCS, which further demonstrates the effectiveness of incorporating label correlation for multi-label active learning. In Figure 2, although there is no clear contribution of label correlation at the beginning of the curves (few example-label pairs are queried), no loss of performance is incurred. As more and more example-label pairs are queried, the label correlation can be more accurately estimated, and thus its contribution becomes evident. (iii) CMAL and CMAL-nC have a similar performance in some cases. The reason is that the datasets used in our experiments only have a small number of examples, and the average number of labels per example is small. As a consequence, the correlation between labels cannot be well estimated, and the contribution of this estimation is not prominent. But due to the contribution of label space sparsity, CMAL-nC can obtain a performance similar to CMAL. This observation motivates us to pursue more reliable label correlation measures in the future.

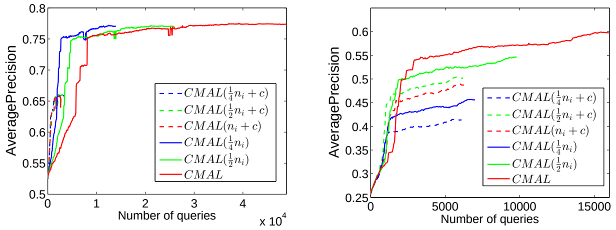
D. Study on Alternative Querying Strategies

In the previous experiments, we strictly followed the multi-instance learning assumption that an example-label pair is deemed as negative if all its subexample-label pairs have been queried and received a negative feedback. Here we consider avoiding querying all subexamples of a negative example-label pair. As we gather more evidence of negative subexample-label pairs for a given example-label pair, the likelihood that the example itself is negative, with respect to the queried label, increases. In addition, if the subexample-level classifier $g_c(\mathbf{x})$ consistently predicts all not-yet-queried subexample-label pairs of the example in question as negative, then the example-label pair is even more likely to be negative. Given these observations, we introduce five variants of CMAL: CMAL($\frac{n_i}{4}$), CMAL($\frac{n_i}{2}$), CMAL($\frac{n_i}{4}$+C), CMAL($\frac{n_i}{2}$+C) and CMAL(n_i+C). In CMAL($\frac{n_i}{4}$), the learner annotates the selected example-label pair as negative if $\sum_{k=1}^{n_{i^*}}[\mathbf{y}_{i^*kc^*} = -1] \geq \lceil \frac{n_{i^*}}{4} \rceil$, namely the first quarter of subexamples of the example are all negative for the target label. In CMAL($\frac{n_i}{4}$+C), the learner annotates the selected example-label pair as negative if $\sum_{k=1}^{n_{i^*}}[\mathbf{y}_{i^*kc^*} = -1] \geq \lceil \frac{n_{i^*}}{4} \rceil$, or $\mathbf{y}_{i^*kc^*} = -1$ for all queried subexample-label pairs and $g_c(\mathbf{x}) < 0$ for all not-queried subexample-label pairs (consensus prediction). The other variants follow the same naming rules. Figure 3 gives the *AveragePrecision* of CMAL and these five variants on two datasets. The end of a curve means all the example-label pairs in the pool are queried.



(a) Birds (b) MSRC-v2

Fig. 2. AveragePrecision of CMAL and its variants.



(a) Birds (b) MSRC-v2

Fig. 3. *AveragePrecision* of CMAL and its five variants. The number of misjudged example-label pairs for the five variants (ordered as in the legend above) are as follows. Birds: 192, 193, 162, 40, 37; MSRC-v2: 266, 243, 245, 231, 124, respectively.

From Figure 3, we observe the following. (i) The five variants stop earlier than CMAL, and they make a smaller number of queries of subexample-label pairs than CMAL. This is because the five variants do not query all subexample-label pairs to select a negative example-label pair, and thus greatly reduce the number of queries, but can misjudge positive example-label pairs as negative. (ii) Although CMAL has a lower performance than its variants at the beginning, it often achieves the best performance as the query budget increases. (iii) The use of the consensus prediction of the subexample-level classifier greatly saves the query cost (i.e., the number of queries for subexample-label pair) of an example-label pair, but results in more misjudged positive example-label pairs and also lower performance. (iv) With or without the use of the consensus of the subexample-level classifier, these variants misjudge more positive example-label pairs as $\theta = \{1, 2, 4\}$ (n_i/θ) increases (decreases). The reason is that as θ increases, the example-label pair can be more quickly deemed as negative and this saves cost, but at the price of increasing the risk of misjudging a positive example-label pair as negative. This leads to a degenerated performance.

Overall, the query cost and performance margin between CMAL and its variants provide options for the user to select an appropriate query strategy for subexample-label pairs based on the budget. For a small budget, the variant with a larger θ is more suitable; otherwise, the variant with a smaller θ should be preferred.

V. Conclusion

We study active learning on multi-label examples that can be segmented (or naturally represented) in multiple subexamples. We observe that the annotators can more easily annotate an example-label pair by annotating subexamples of the target example. Based on this observation, we propose a cost-effective multi-label active learning approach called CMAL. CMAL first selects the most uncertain example-label pair and then queries its most likely positive subexample-label pair. Experimental results on multi-instance datasets demonstrate that CMAL is able to achieve higher accuracy than other related methods, and at an inferior cost. The code of CMAL is available at http://mlda.swu.edu.cn/codes.php?name=CMAL.

The assumption that all subexamples of negative examples are negative may not hold true in some domains. For example, in document classification, some paragraphs may be negative for the topic of the document, but some other paragraphs may be postive. In addition, the assumption that the query cost of all subexample-label pairs is equal to that of the example-label pair is too optimistic. These two assumptions will be relaxed in our future work.

Acknowledgment

The authors thank the anonymous reviewers for their helpful comments, which contributed to the improvement of the paper. This work is in part supported by the Natural Science Foundation of China (61872300, 61741217 and 61873214), the Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228 and cstc2016jcyjA0351), the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2017A05), and the Chongqing Graduate Student Research Innovation Project (No. CYS18089).

References

[1] X. Li and Y. Guo, "Active learning with multi-label svm classification," in *IJCAI*, 2013, pp. 1479–1485.

[2] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor, "Active learning for sparse bayesian multilabel classification," in *KDD*, 2014, pp. 472–481.

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

[4] S.-J. Huang and Z.-H. Zhou, "Active query driven by uncertainty and diversity for incremental multi-label learning," in *ICDM*, 2013, pp. 1079–1084.

[5] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *TPAMI*, vol. 36, no. 10, pp. 1936–1949, 2014.

[6] W. Liu and I. W. Tsang, "Making decision trees feasible in ultrahigh feature and label dimensions," *JMLR*, vol. 18, no. 1, pp. 2814–2849, 2017.

[7] Y.-F. Li, J.-H. Hu, Y. Jiang, and Z.-H. Zhou, "Towards discovering what patterns trigger what labels," in *AAAI*, 2012, pp. 1012–1018.

[8] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *TKDE*, vol. 26, no. 8, pp. 1819–1837, 2014.

[9] W. Liu, I. W. Tsang, and K.-R. Müller, "An easy-to-hard learning paradigm for multiple classes and multiple labels," *JMLR*, vol. 18, no. 1, pp. 3300–3337, 2017.

[10] R. Retz and F. Schwenker, "Active multi-instance multi-label learning," in *Analysis of Large and Complex Data*, 2016, pp. 91–101.

[11] S.-J. Huang, N. Gao, and S. Chen, "Multi-instance multi-label active learning," in *IJCAI*, 2017, pp. 1879–1892.

[12] P. Donmez and J. G. Carbonell, "Proactive learning: cost-sensitive active learning with multiple imperfect oracles," in *CIKM*, 2008, pp. 619–628.

[13] C. Zhang and K. Chaudhuri, "Active learning from weak and strong labelers," in *NIPS*, 2015, pp. 703–711.

[14] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, "Multi-label active learning from crowds," *arXiv preprint arXiv:1508.00722*, 2015.

[15] S. J. Huang, J. L. Chen, X. Mu, Z. H. Zhou, S. J. Huang, J. L. Chen, X. Mu, and Z. H. Zhou, "Cost-effective active learning from diverse labelers," in *IJCAI*, 2017, pp. 1879–1885.

[16] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *TCSVT*, vol. 27, no. 12, pp. 2591–2600, 2017.

[17] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *TPAMI*, vol. 40, no. 1, pp. 7–19, 2017.

[18] S.-J. Huang, S. Chen, and Z.-H. Zhou, "Multi-label active learning: Query type matters," in *IJCAI*, 2015, pp. 946–952.

[19] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[20] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *arXiv preprint arXiv:1612.03365*, 2016.

[21] M.-A. Carbonneau, E. Granger, and G. Gagnon, "Bag-level aggregation for multiple instance active learning in instance classification problems," *arXiv preprint arXiv:1710.02584*, 2017.

[22] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *NIPS*, 2008, pp. 1289–1296.

[23] J. Meessen, X. Desurmont, J.-F. Delaigle, C. De Vleeschouwer, and B. Macq, "Progressive learning for interactive surveillance scenes retrieval," in *CVPR*, 2007, pp. 1–8.

[24] D. Zhang, F. Wang, Z. Shi, and C. Zhang, "Interactive localized content based image retrieval with multiple-instance active learning," *Pattern Recognition*, vol. 43, no. 2, pp. 478–484, 2010.

[25] J. Melendez, B. van Ginneken, P. Maduskar, R. H. Philipsen, H. Ayles, and C. I. Sánchez, "On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis," *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 1013–1024, 2016.

[26] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2003, pp. 577–584.

[27] B. Settles, "Active learning literature survey," *Technical report 1648, University of Wisconsin-Madison*, 2009.

[28] N. Youngs, D. Penfoldbrown, K. Drew, D. Shasha, and R. Bonneau, "Parametric bayesian priors and better choice of negative examples improve protein function prediction," *Bioinformatics*, vol. 29, no. 9, p. 1190, 2013.

[29] Q. Tan, Y. Yu, G. Yu, and J. Wang, "Semi-supervised multi-label classification using incomplete label information," *Neurocomputing*, vol. 260, pp. 192–202, 2017.

[30] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for miml instance annotation," in *KDD*, 2012, pp. 534–542.

[31] C. C. Chang and C. J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.