Multi-Label Active Crowdsourcing Learning with Uncertainty Measure (Technical Report)

Ming Wu, Jing Zhang
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing 210094, China
{mingwu, jzhang}@njust.edu.cn

Abstract

Traditional supervised machine learning obtains labels from high-quality oracles, which is high cost and time-consuming and does not consider security. Since multi-label active learning becomes a hot topic, it is more challenging to train efficient classification models and reduce the label cost. To deal with this problem, in this paper, we work on the multi-label active learning using an economic and efficient strategy—crowdsourcing, where querying the labels from multiple low-cost annotators with various expertise on crowdsourcing platforms instead of resorting to a high-quality oracle. To eliminating the effects of annotation noises due to imperfect annotators, we propose the Multi-Label Active Crowdsourcing Learning with Uncertainty measure (MACLU) to build a probabilistic model to compute the annotation consensus and estimate the parameters of the classifier simultaneously, which also introduces the similarity information of instances into consideration. Then, an instance-label-annotator triplets selection strategy is proposed to actively choose the most informative instances and labels, and the most reliable annotators. Experimental results on two real-world data sets show that the performance of MACLU is superior to existing methods.

1 Introduction

Due to the multiplicity of instances in the real world, multi-label learning has already attracted wide attention in the field of machine learning (Zhang and Zhou, 2007, 2013; Zhou and Zhang, 2017). In traditional classification tasks, each instance only has one label to represent its category (Jing et al., 2018). However, it is common that multiple labels belong to an instance simultaneously in many practical tasks. As shown in Fig. 1(a), the image has 2 labels *sea* and *sunset* at the same time, while Fig. 1(b) has 3 labels (*sea, mountain and tree*). In comparison with single-label classification tasks, the number of labels in multi-label learning tasks increases exponentially, which leads to significant growth of cost since each possible label will be checked whether it belongs to the instance. To reduce the cost of multi-label learning, actively selecting the most valuable instances for training becomes a widely used approach (Singh et al., 2009; Yang et al., 2009).

Traditional multi-label active learning achieves annotations from reliable and expensive experts, based on the assumption that the annotations provided by experts are correct. With the rapid rise of crowdsourcing platforms, a large number of annotations can be obtained fast at a low cost on the platforms. In the process of obtaining annotations by crowdsourcing, multiple low-quality annotations will be provided by different imperfect annotators for each instance, and then an integrated annotation will be inferred from the set of noisy annotations since Snow et al. (2008) has discussed the effectiveness of collecting data annotations from multiple non-expert annotators instead of resorting to experts for the ground truth. Snow et al. (2008) also illustrates that a well-designed

Preprint. Under review.

(a) sea+sunset

(b) sea+mountain+tree

(c) desert

(d) foliage+tree

Figure 1: Examples of multi-label images.

annotation consensus method can make a significant improvement of data quality. Previous works about crowdsourcing mainly focus on annotation consensus (Zhang and Zhou, 2013), and a majority of annotation consensus methods only study on the single-label scenario (Dawid and Skene, 1979; Whitehill et al., 2009; Zhang et al., 2014; Wu et al., 2017).

We study the problem of secure multi-label active learning tasks to reduce the cost of model training while achieving high-quality performance. Since the high requirement of the number of annotations inevitably in multi-label tasks, it is still expensive and time-consuming to obtain annotations provided by experts. To address this problem, we take the wisdom of crowds into consideration, which is a more economic approach to collect the annotations. Some works think of the multi-label task as multiple independent single-label tasks (Yan et al., 2011). However, they ignored that label correlation is one of the significant factors to affect learning performance. As shown in Fig. 1(c) and Fig. 1(d), when an image has a label *desert*, it is almost impossible for the image to have another label *sea*, meanwhile, the label *foliage* usually come with *tree* in the same image. Since the label correlation is helpful for the model learning, we propose a novel Multi-Label Active Crowdsourcing Learning with Uncertainty measure (MACLU), which introduces the information of label correlation to enhance the features of instances and estimates the multi-label classifier and reliability of annotators simultaneously in a probabilistic model. Then an instance-label-annotator triplets selection strategy is applied to select the most reliable annotator to label the most informative instance and label. Moreover, compared with the previous works, the proposed method focuses on security and privacy protection, since it can select reliable annotators without prior information and train a robust classifier. The main contributions of this study are summarized as follows:

- A novel MACLU method is proposed to exploit crowdsourcing in the scenario of multi-label active learning tasks. MACLU utilizes the information of label correlation to estimate the annotation consensus and reliability of annotators.
- We apply a novel active learning strategy to reduce the cost, due to multi-label tasks are expensive and time-consuming traditionally. We take the correlation of labels into consideration, and the most valuable instance-label-annotator triplets are selected to make an improvement.
- Experimental results demonstrate that our proposed approach outperforms baselines and state-of-the-art methods.

2 Related Work

2.1 Active Learning from Crowds

Active learning is a strong tool to actively select the most informative data and achieve the adequate accuracy of the trained model with a limited budget(Malago et al., 2014). Due to the well performance of active learning for making a trade-off between cost and accuracy of the trained model, active learning strategies has been studied in many applications, such as text classification (Hoi et al., 2006), video retrieval (Hauptmann et al., 2006) and image categorization and retrieval (Zhang and Chen, 2002), and so on(Cui et al., 2020a,b). Previous works mainly focus on selecting most valuable instances for labeling, and the selection criterions can be generally classified into three categories: 1) Uncertainty sampling (Lewis and Gale, 1994); 2) Expected Model Change (Settles, 2009); 3) Ensemble methods (Huang and Zhou, 2013; Jing et al., 2015). Furthermore, when we consider the active learning framework in the crowdsourcing scenario, we not only need to select the most valuable instance, but also assign the best annotator for the labeling task. Previous works mainly try to design the criterion for instance and annotator selections, for example, Yan et al. (2011) builds a probabilistic model to estimate the reliability of annotators and then provides a criterion to select instance-annotator pairs, and Huang and Zhou (2013) presents a novel instance-label pairs query criterion, which takes advantage of uncertainty and diversity in the data space. Some works focus on learning the knowledge of annotators to aggregate the noisy annotations, e.g. Zhong et al. (2015) allows annotators to provide 'unsure' option to reduce the error rate of labeling tasks, and then modeling the annotators' reliability, Fang et al. (2014) utilizes a transfer learning method to help compute the reliability of annotators from auxiliary domains, and Song et al. (2018) obtain the annotations and the confidence from annotators simultaneously, then makes annotation consensus with the confidential information. However, these methods cannot be applied in multi-label tasks.

2.2 Multi-Label Learning from Crowds

Multi-label learning considers the tasks where an instance is associated with multiple labels, which is common in real life (Zhao and Guo, 2015; Wu et al., 2014; Boutell et al., 2004). Thus, multi-label learning from crowdsourcing becomes a hot topic for the past few years. Most researches focus on annotation consensus approaches for multi-label tasks, for example, Li et al. (2018) models the label correlations in a probabilistic model, which is an extension of single-label aggregation methods, and Zhang and Wu (2018, 2019) present an EM-based algorithm to estimate the true labels of each instance as well as the reliability of each annotator.

2.3 Multi-Label Active Learning from Crowds

In addition, in the traditional multi-label active learning research, strategies for the instance and label selection are two critical topics. For example, Ye et al. (2015) utilizes the combination of instance feature and label correlation information to select instance-label pairs, Wu et al. (2014) uses the whole uncertainty of each label as the selection criterion, Reyes et al. (2018) proposes an instance selection strategy based on instance predictions and label inconsistency, while Li and Guo (2013) and Yang et al. (2009) also considers the uncertainty strategy with some other influence factors for SVM classification problems. However, in the research of multi-label active learning from crowds, selecting the most reliable annotator should also be carefully discussed.

To the best of our knowledge, there are a few articles research on multi-label active learning in a crowdsourcing scenario, especially in the case of high-security requirements. Yu et al. (2020) provides an annotation consensus method for multi-label tasks actively, whereas it does not present a learning model for categorization. Li et al. (2015) proposes a MAC method to build a probabilistic model with a simple logistic regression function. Our proposed MACLU jointly models the classifier and annotators' reliability in a probabilistic model, and actively selects instance-label-annotation triplets for reducing cost.

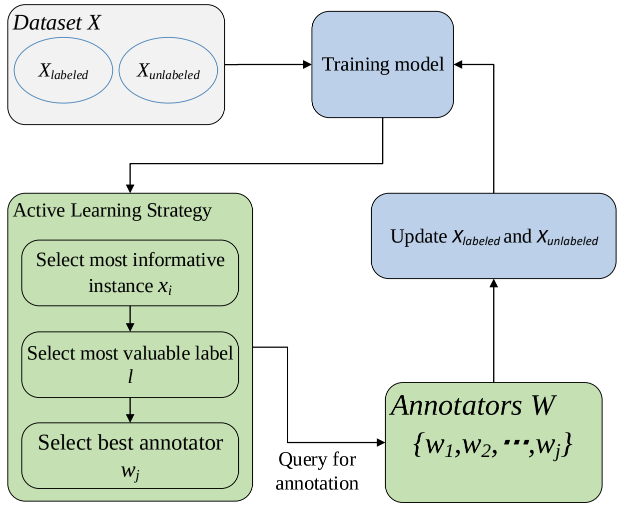


Figure 2: A general framework of the proposed MACLU method.

3 The Proposed Method

3.1 Preliminaries

In this paper, we focus on the pool-based multi-label active learning from crowds, so the whole data set is defined as $\mathcal{X}$, which comprises two parts as follows, $\mathcal{X}_L$ denotes the labeled training data that $\mathcal{X}_L = \{(x_i, y_i, \{\hat{y}_i\})\}_{i=1}^{D_L}$ and $\mathcal{X}_U$ denotes the unlabeled training data that $\mathcal{X}_U = \{x_i\}_{i=D_L+1}^{D_L+D_U}$, where D_L and D_U denotes the number of labeled and unlabeled data separately. Commonly, $\mathcal{X}_L$ forms a tiny part of $\mathcal{X}$. Each instance x_i is an M dimension features where $i \in \{1, ..., I\}$. y_i denotes the ground truth labels of instance x_i which is a K dimension vector as $y_i = \{-1, 1\}^{1\times K}$, when the l_{th} element $(l \in \{1, ..., K\})$ of y_i is 1 means the instance x_i has the l_{th} label, and -1 means the instance don't have the corresponding label. $\hat{y}$ represents the noisy annotations obtained from crowdsourcing, where $\hat{y}_i = \{\hat{y}_{ij}^l\}_{j=1,l=1}^{J,K}, j \in \{1, ..., J\}$. We assume a set of J annotators $\{w_j\}^J$, so that $\hat{y}_{ij}^l$ represents the annotation provided by annotator w_j to the instance x_i on the l_{th} label, and $\hat{y}_{ij}^l = \{-1, 0, 1\}$, where $\hat{y}_{ij}^l = 1(-1)$ indicates annotator w_j gives a positive(negative) annotation to instance x_i on the l_{th} label, and $\hat{y}_{ij}^l = 0$ means the annotator w_j gives no annotation to the instance. Traditionally, all annotators need to provide annotations once the instance is selected, however, in this paper, we do not need all annotators label the selected instance. The goal of our proposed method is training a probabilistic model to estimate the annotation consensus and annotators' reliability, meanwhile, a novel active selection strategy is utilized to select the most informative instance-label-annotator triplet to improve the performance of the training model and reduce the cost.

3.2 Annotation Consensus

Fig. 2 is a general framework of MACLU, which comprises two steps, model learning, and active selection. Since the annotations are provided by non-expert annotators on crowdsourcing platforms, it is not hard to understand that the quality of the annotations is various due to the different expertise and dedication of the annotators. Typically, in the field of crowdsourcing, repeated labeling each instance and obtaining the integrated label as appropriate substitutions for the unknown ground truth is successfully used in multiple applications (Zhang et al., 2015; Demartini et al., 2012). Thus, a well-designed annotation consensus algorithm is helpful in the context of crowdsourcing. However, most of the annotation consensus methods currently only care about the annotations and do not consider the instances. In this paper, we introduce the features of instances to build a probabilistic model, so that label correlation information can help the model training. The target of our inference model is to achieve a minimum of the empirical risk, so that we can use the integrated annotation $\hat{y}_i^l$ as the ground truth y_i^l:

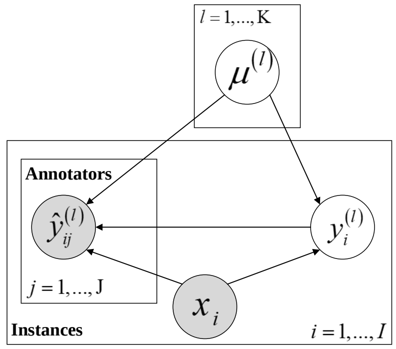


Figure 3: Probabilistic graphical model of the proposed MACLU method

$$\mathcal{E} = \frac{1}{I \cdot K} \sum_{i}^{I} \sum_{l}^{K} \mathbb{I}(\hat{y}_i^l \neq y_i^l), \tag{1}$$

where $\hat{y}_i^l$ denotes the inferred $l-th$ label of instance x_i from the noisy annotation set $\{\hat{y}_{ij}^l\}$, and $\mathbb{I}(*)$ is an indicator function, whose output equals 1 if the input is true, or else the output will be 0.

Fig. 3 illustrates the probabilistic graphical model representation of the proposed MACLU. In this model, the observed variable $\hat{y}_{ij}^l$ depends on both the features and the ground truth of instance x_i. Meanwhile, the latent variable y_i^l, which is the ground truth of instance x_i, is affected by the features of instance x_i. So the conditional probabilistic distribution is represented as

$$P(\hat{y}_{ij}^l, y_i^l | x_i, \mu^l) = P(\hat{y}_{ij}^l | x_i, y_i^l, \mu^l) P(y_i^l | x_i, \mu^l). \tag{2}$$

Furthermore, the joint probability distribution of the whole training data is built as follows:

$$P(\{\hat{y}_{ij}^l\}, \{y_i^l\} | \{x_i\}, \{\mu^l\}) = \prod_i \prod_j \prod_l P(\hat{y}_{ij}^l | x_i, y_i^l, \mu^l) \prod_i \prod_l P(y_i^l | x_i, \mu^l) \tag{3}$$

To estimate the ground truth of instances, we train a set of classifiers $\{h^{(l)}\}$ from the training set to correctly categorize instances on each label. Then we purpose to minimize the overall error rate of the classifiers $\{h^{(l)}\}$. Typically, the classifiers are trained separately on each label, however, the correlation of labels is usually not taken into account. Here we generate the enhanced features $\{\hat{x}_i\}$ for each instance with the complementary label correlation information, which has a significant benefit on model learning. For the problem of multi-label learning, We assume that 1) the instances similar in the feature space should also have similar label distribution, 2) the closer the neighborhood is, the more influence it has on the instance. Therefore, we exploit the weighted mean of the top-k nearest neighbors to construct the enhanced features of the instances. For each instance x_i, its enhanced feature is represented as $\hat{x}_i = [x_i; \psi_i]$, where ψ_i is the complementary information of x_i. For convenience, we still use x_i to represent the enhanced feature in the later sections. ψ_i is a K dimension vector to reflect the label correlation information,

$$\psi_i = \sum_{x_n \in \mathcal{N}_k(x_i)} (wt_n \cdot \hat{y}_n), \tag{4}$$

where $\mathcal{N}_k(x_i)$ is a function to output the top-k nearest neighbors of x_i, wt_n is the weight of instance x_n. For ease of understanding, we define $x_i^{(d)}$ the d-th nearest neighbor of x_i ($1 \leq d \leq k$), i.e.,

$$wt_d = \mathcal{D}(x_i, x_i^{(k-d+1)}) / \sum\nolimits_{x_i^{(d)} \in \mathcal{N}_k(x_i)} \mathcal{D}(x_i, x_i^{(d)}), \tag{5}$$

where $\mathcal{D}(a, b)$ denotes the distance between points a and b.

Since the features of instances are similar on the complementary dimensions $\{\psi_i\}$, to better deal with the classification problem, we use SVM to implement the classifiers $\{h^{(l)}\}$, i.e., $h^{(l)}(x_i) = w^{(l)}\sigma(x_i)$, where $\sigma()$ is the kernel function, i.e.,

$$p(y_i^{(l)} | \mu, x_i) = \phi(h^{(l)}(\hat{x}_i))^{\mathbb{I}(y_i^{(l)}=1)} (1 - \phi(h^{(l)}(\hat{x}_i)))^{\mathbb{I}(y_i^{(l)}=-1)}, \tag{6}$$

where the function $\phi(*)$ denotes the *sigmoid function* formulated as $\phi(t) = 1/(1 + exp(-t))$.

Meanwhile, for each annotator, we assign a latent variable $r_j^{(l)}$ to indicate the reliability of annotator w_j on label l, where $r_j^{(l)} \in (0, 1)$. For example, when $r_j^{(l)}$ trends to 0, the annotator w_j may be a malicious adversary who always provide an incorrect answer, also, the annotator w_j may be an expert if $r_j^{(l)}$ trends to 1, and $r_j^{(l)} = 0.5$ means the annotator is in a wild guess to provide an annotation. $r_j^{(l)}$ also represents the probability that instance w_j gives a true annotation on label l. Due to the reliability of the annotator depends on specific instance and label, we define it as a *logistic regression model* where we use the *sigmoid function* as the link function,

$$r_j^{(l)} = \phi(\mathcal{L}_{\pi_j^{(l)}}(x_i)), \tag{7}$$

where $\pi_j^{(l)}$ denotes the parameter of $r_j^{(l)}$ and $\mathcal{L}_{\pi_j^{(l)}}(*)$ denotes the *logistic regression model*, so that,

$$p(\hat{y}_{ij}^l = y_i^l | r_j^{(l)}, x_i) = \phi(\pi_j^{(l)} x_i). \tag{8}$$

Therefore, the parameters represent as $\mu = \{w^{(l)}, \pi_j^{(l)}\}$. To estimate the parameters and the ground truth of the instances, we use an EM-based algorithm (Dempster, 1977) to maximize the likelihood.

In *E-step*, we calculate the posterior probability of ground truth $\{y_i\}$ given the parameter μ from last M-step, which gives:

$$\begin{aligned} Pr(y_i^l) = P(y_i^l | \{\hat{y}_{ij}^l\}, \mu, x_i) \propto P(y_i^l | \{\hat{y}_{ij}^l\}, x_i, w^{(l)} \{\pi_j^{(l)}\}) \\ \propto P(y_i^l | w^{(l)}, \hat{x}_i) P(\{\hat{y}_{ij}^l\} | x_i, \{\pi_j^{(l)}\}, y_i^l). \end{aligned} \tag{9}$$

According to Eq. 6 and Eq. 8, we have

$$\begin{aligned} P(y_i^l = 1) = \phi(h^{(l)}(\hat{x}_i)) \prod_{y_{ij}^l} \{\phi(\pi_j^{(l)} x_i)^{\mathbb{I}(y_{ij}^l = 1)} \phi(-\pi_j^{(l)} x_i)^{\mathbb{I}(y_{ij}^l = -1)}\}, \\ P(y_i^l = -1) = (1 - \phi(h^{(l)}(\hat{x}_i))) \prod_{y_{ij}^l} \{\phi(\pi_j^{(l)} x_i)^{\mathbb{I}(y_{ij}^l = -1)} \phi(-\pi_j^{(l)} x_i)^{\mathbb{I}(y_{ij}^l = 1)}\} \end{aligned} \tag{10}$$

In *M-step*, we update the parameters by optimizing a complementary function $\mathcal{Q}(\mu, \mu^{old})$ defined as:

$$\begin{aligned} \mathcal{Q}(\mu, \mu^{old}) &= \mathbb{E}_{y_i^l | \hat{y}_{ij}^l, \mu^{old}} [lnP(\hat{y}_{ij}^l, y_i^l | \mu, x_i)] \\ &= \mathbb{E}[lnP(\hat{y}_{ij}^l | x_i, y_i^l, \mu^l) P(y_i^l | x_i, \mu^l)]. \end{aligned} \tag{11}$$

The updated parameters μ^{new} are computed as:

$$\mu^{new} = \arg\max_{\mu} \mathcal{Q}(\mu, \mu^{old}). \tag{12}$$

Then we have:

$$\begin{aligned} w^{(l)} &= w^{(l)} + \nabla w^{(l)} \\ \nabla w^{(l)} &= -\eta_w \frac{\partial \mathcal{Q}(\mu)}{\partial w^{(l)}} \\ \frac{\partial \mathcal{Q}(\mu)}{\partial w^{(l)}} &= -\sum_i [y_i^{(l)} - \phi(w^{(l)} x_i)] x_i \end{aligned} \tag{13}$$

$$\begin{aligned} \pi_j^{(l)} &= \pi_j^{(l)} + \nabla \pi_j^{(l)} \\ \nabla \pi_j^{(l)} &= -\eta_\pi \frac{\partial \mathcal{Q}(\mu)}{\partial \pi_j^{(l)}} \\ \frac{\partial \mathcal{Q}(\mu)}{\partial \pi_j^{(l)}} &= -\sum_i [y_i^{(l)} - \phi(w^{(l)} x_i y_{ij}^l)] x_i y_{ij}^l, \end{aligned} \tag{14}$$

where η_w and η_π are the learning rates.

After convergence, the integrated annotations of the instances will be inferred by the updated parameters $\{\mu^{(l)}\}$. That is,

$$\hat{y}_i^l = \arg\max P(y_i^l | \{\hat{y}_{ij}^l\}, \mu, x_i) \tag{15}$$

3.3 Active Learning Strategies

To choose the most valuable instance for the above model, we employ a novel active learning strategy, which considers the diversity and uncertainty of instances. After obtaining the instance, we also select the appropriate label and ask the most reliable annotator to label it.

3.3.1 Instance Selection

To design the instance selection strategy, we utilize a combination measure of uncertainty and diversity for each instance. That is:

$$S(x_i) = \frac{CI(x_i)}{U(x_i)} \tag{16}$$

$CI(x_i)$ is an extension of the commonly used information measure LCI (Label Cardinality Inconsistency), which combined with the number of times the instance was queried as follows:

$$CI(x_i) = \frac{|\sum_l \mathbb{I}(\hat{y}_i^l = 1) - \frac{1}{D_L}\sum_i^{D_L}\sum_l \mathbb{I}(y_i^l = 1)|}{max\{\epsilon, card(x_i)\}}, \tag{17}$$

where $\hat{y}_i^l$ is estimated by Eq. 15, $\epsilon \in (0,1)$ is a constant to avoid the denominator to be 0. $card(x_i)$ denotes the number of annotations x_i already has. Thus, $CI(x_i)$ tends to select the instance with the maximum LCI and least annotations.

$U(x_i)$ measures the uncertainty of instance x_i. That is:

$$U(x_i) = \frac{1}{K}\sum_l |h^{(l)}(x_i) - 0.5|, \tag{18}$$

where $h^{(l)}$ is the currently trained classifiers on training data. The smaller $U(x_i)$ is, the greater the uncertainty of x_i is.

Therefore, we select the most valuable instance x_i^* as follows:

$$x_i^* = \arg\max_i S(x_i). \tag{19}$$

To avoid that the same instance is selected many times, we set a constraint that when each label of an instance is annotated more 5 times, we will remove this instance from unlabeled training dataset. Also, annotators with low reliability will have the chance to practice by increasing their weights when reliable annotators have finished the annotation tasks.

3.3.2 Label Selection

Once the instance x_i^* is selected, we select the appreciate label of it. Here we choose the label with the uncertainty information. That is:

$$l^* = \arg\max_l (0.5 - |h^{(l)}(x_i) - 0.5|). \tag{20}$$

3.3.3 Annotator Selection

Annotators have various reliabilities on different labels. After we select the instance-label pair, the annotation should be queried from the most reliable annotator. The reliability of annotators are estimated in the current model. Therefore, the annotator selection is defined as:

$$\begin{aligned} w_j^* &= \arg\max_j r_j^{(l)} \\ &= \arg\max_j \phi(\pi_j^{(l)} x_i) \end{aligned} \tag{21}$$

Since the reliability of annotators changes on different instances, this strategy can avoid the situation that the same annotator always be selected to labeling. After selecting the instance-label-annotator triplets, new annotations add to the training data, and the model will be updated using these high quality annotations.

Algorithm 1: The MACLU Algorithm

Input: Dataset set $\mathcal{X}$, parameters μ

Output: Classifiers $\{h^{(l)}\}$, reliabilities $\{r_j^{(l)}\}$

Initialization: Divide the data set into three parts randomly, e.g. labeled training data $\mathcal{X}_L$, unlabeled training data $\mathcal{X}_U$ and test data $\mathcal{X}_T$. Calculate the complementary information ψ of each instance by Eq. 4. Initialize the parameters μ using *scikit-learn* on noisy labeled data $\mathcal{X}_L$;

while *model can be improved and $\mathcal{X}_U$ is not empty* **do**

    Active Learning Selection;

    Select instance x_i^* by Eq. 19;

    Select label l^* by Eq. 20;

    Select annotator w_j^* by Eq. 21;

    Query annotation $\hat{y}_{i^*j^*}^{l*}$ from annotator w_j^*;

    Add the selected instance-label-annotator triplet to labeled data $\mathcal{X}_L$;

    Model Training;

    while *not convengence* **do**

        Perform *E-step* by Eq. 10;

        Perform *M-step* by Eq. 13 and Eq. 14;

    end

    Get the updated classifiers $\{h^{(l)}\}$ and parameters μ

end

3.3.4 Pseudo-Code of the Algorithm

The overall algorithm of MACLU is summarized in Algorithm. 1. First, an initialization step is implemented. Due to the inherent defect of EM-based algorithms that is sensitive to initial parameters, we employ a spectral clustering algorithm on the training data firstly to avoid convergence to a local optimum, then we find the k-nearest neighbors of each instance in the same cluster. Secondly, the active learning strategy and model training procedure run iteratively until the model can not be improved or the unlabeled data set is empty. Finally, we can get the classifiers $\{h^{(l)}\}$ and the parameters of each annotator. When a new instance is obtained, we can first compute its enhanced feature, and predict its labels by the classfiers $\{h^{(l)}\}$.

4 Experiments

4.1 Experimental Setup

Our approach is implemented with the *scikit-learn* module, based on python. Two real-world data sets are used in our experiments as shown in Table. 2. The data set *Image* (Boutell et al., 2004) contains 2407 multi-label images and an output space of 6 labels, and over 7% of the images are assigned more than one label. Data set *Scene* Zhang and Zhou (2007) contains 2000 natural scene images and 5 possible labels, and the number of images which have more than one label is over 22% of the data set. The details of these two datasets are shown in Table. 1.

In the experiments, we set the parameter k as 5, and ϵ is set as 0.5. For each dataset, we split them into three parts, e.g. labeled training data, unlabeled training data and test data, which are separately 5%, 50% and 45% of the whole data set $\mathcal{X}$. Particularly, in the labeled data, the instances, labels and annotators are all selected randomly. Then we simulate 5 annotators as the crowds to provide annotations. Moreover, we generate a set of confusion matrices for each annotator which represent the ability of the annotator. The confusion matrix is built as $M^{C\times C}$, where C means the number of categories in each label, for binary classification we have $C = 2$, the entry m_{pq} of M means the probability that the annotator provides annotation p to the instance whose ground truth is q, and the diagonal entries of the matrix are randomly extracted from a uniform distribution $U(0.5, 0.9)$. The setting of parameters suits the actual investigation, that most of the annotators have the accuracy higher than 0.5 except malicious ones, but lower than experts. When an annotator is queried to give an annotation, we generate the annotation using the confusion matrix. On each data set, we run the methods 3 times to calculate the average performance.

Table 1: DETAILS OF TWO REAL-WORLD DATA SETS USED IN THE EXPERIMENTS

DATASETS	LBS.	#INS	CARD.	DENS.
Image	beach	427	1.07	0.18
	field	433		
	foliage	397		
	mountain	533		
	sunset	364		
	urban	431		
Scene	desert	409	1.24	0.25
	mountains	458		
	sea	580		
	sunset	465		
	trees	560		

Note: LBS. shows the list of possible output labels and #INS represents the corresponding number of instances, CARD. represents the *cardinality* of the data set while DENS. means the *density* of the data set.

Table 2: STATISTICS OF TWO REAL-WORLD DATA SETS USED IN THE EXPERIMENTS

DATASETS	#FEA	POSS LBS.	#LBS	#INS	TOTAL
Image	294	{*beach, field, foliage, mountain, sunset, urban*}	1	2230	2407
			2	176	
			3	1	
Scene	294	{*desert, mountains, sea, sunset, trees*}	1	1543	2000
			2	442	
			3	15	

Note: #FEA represents the number of features; POSS LBS. represents the set of possible labels on each instance; #LBS and #INS represents the number of labels on each instance and corresponding number of instances respectively.

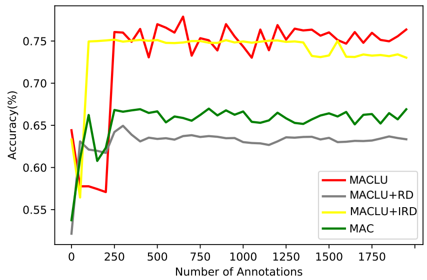
Accuracy is a typical measure metric in classification evaluation, which is defined as the ratio of correct predictions among the total number of testing instances. In this paper, we use average *Accuracy* as one of the evaluation metric. The second measure metric we used in the experiments is F-measure (F_1 score), which is the harmonic mean of the precision and recall, defined as $F_1 = 2/(recall^{-1} + precision^{-1})$.

To the best of our knowledge, there are few algorithms study on the problem of multi-label active learning from crowds. Therefore, we choose one alternative approach and two other baselines to conduct comparisons.

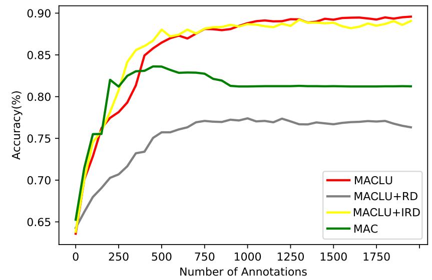
- *MACLU_IRD* This baseline trains the proposed model in this paper, while in active learning step, we randomly select the instance to query, then select the label and annotator by our proposed active learning strategies.
- *MACLU_RD* In this approach, the proposed model is trained, while in active learning step, we randomly select the instance-label-annotator triplet to query.
- *MAC* To the best of our knowledge, MAC Li et al. (2015) is the first approach to study the problem of active multi-label learning from crowds.

4.2 Experimental Results

As shown in Fig. 4, our proposed MACLU outperforms other comparison methods in terms of *Accuracy*. We find that: 1) MACLU converges faster than others so that it can reduce the cost effectively, which less than 500 labels. 2) Compared with random selections (*RD* and *IRD*), our proposed model improves the accuracy significantly. 3) From Fig. 4(b), we find that our proposed model do the annotation consensus better, so that after convergence, the accuracy of MACLU is higher than others. 4) Due to the data imbalance, i.e. in data set *Image*, the number of negative labels



(a) Image

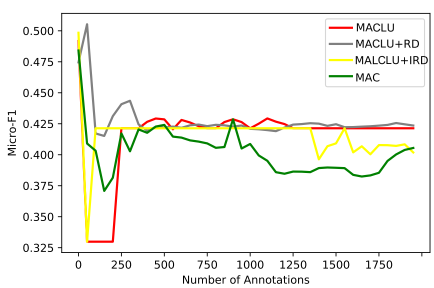


(b) Scene

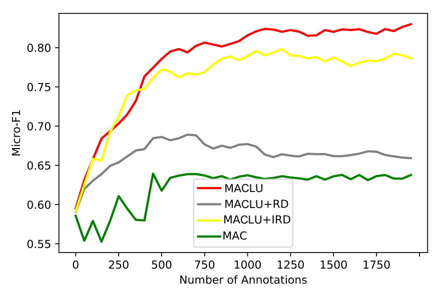
Figure 4: Comparison results of three methods in terms of *Accuracy*.

is much larger than positive ones, the performances of the methods have larger fluctuations. Since the *Accuracy* can not well reflect the performances in the case of imbalanced data, we measure the methods in terms of *F1-micro*.

The comparison results in terms of *F1-micro* are shown in Fig. 5, we still see that our proposed method has the best performance and converges fast. In the case of imbalanced data, i.e. on data set *Image*, the performance of our method is stable.



(a) Image



(b) Scene

Figure 5: Comparison results of three methods in terms of *F1-mirco*.

5 Conclusion

We focus on the secure multi-label active learning problem in the context of crowdsourcing, so that when we need to query a label, instead of querying from an expert oracle, multiple non-expert annotators are available to provide a set of noisy annotations. Multi-label active learning from crowds can make a trade-off between model cost and accuracy. However, it is challenging to obtain high-quality integrated labels from noisy annotations. To address this problem, we propose a novel MACLU approach. The proposed MACLU builds a probabilistic model to estimate the annotation consensus and the reliability of annotators, which takes the label correlation into consideration. Meanwhile, an active learning paradigm is also used to reduce the cost. MACLU comprises an active learning strategy to select instance-label-annotator triplets that are most helpful for model training. Experimental results on two real-world data sets show that the proposed MACLU approach outperforms baselines and state-of-the-art methods.

Acknowledgment

This work was sponsored the National Natural Science Foundation of China under grants 62076130 and the Open Research Projects of Zhejiang Lab (No. 2019KD0AD01/015).

References

Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. 2004. Learning multi-label scene classification. *Pattern recognition* 37, 9 (2004), 1757–1771.

Shicheng Cui, Qianmu Li, and Shu-Ching Chen. 2020a. An adversarial learning approach for discovering social relations in human-centered information networks. *EURASIP Journal on Wireless Communications and Networking* 2020, 1 (2020), 1–19.

Shicheng Cui, Tao Li, Shu-Ching Chen, Mei-Ling Shyu, Qianmu Li, and Hong Zhang. 2020b. DISL: deep isomorphic substructure learning for network representations. *Knowledge-Based Systems* 189 (2020), 105086.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*. 469–478.

A. P. Dempster. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39 (1977).

Meng Fang, Jie Yin, and Dacheng Tao. 2014. Active learning for crowdsourcing using knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

Alexander G Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang, and Ming-Yu Chen. 2006. Extreme video retrieval: joint maximization of human and computer performance. In *Proceedings of the 14th ACM international conference on Multimedia*. 385–394.

Steven CH Hoi, Rong Jin, and Michael R Lyu. 2006. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*. 633–642.

Sheng-Jun Huang and Zhi-Hua Zhou. 2013. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th international conference on data mining*. IEEE, 1079–1084.

Jing, Zhang, Victor, S, Sheng, Tao, Li, Xindong, and Wu. 2018. Improving Crowdsourced Label Quality Using Noise Correction. *IEEE Transactions on Neural Networks & Learning Systems* 5 (2018), 1675–1688.

Z. Jing, X. Wu, and V. S. Sheng. 2015. Active Learning With Imbalanced Multiple Noisy Labeling. *IEEE Transactions on Cybernetics* 45, 5 (2015), 1081–1093.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*. Springer, 3–12.

Shao-Yuan Li, Yuan Jiang, Nitesh V Chawla, and Zhi-Hua Zhou. 2018. Multi-label learning from crowds. *IEEE Transactions on Knowledge and Data Engineering* 31, 7 (2018), 1369–1382.

Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. 2015. Multi-label active learning from crowds. *arXiv preprint arXiv:1508.00722* (2015).

Xin Li and Yuhong Guo. 2013. Active Learning with Multi-Label SVM Classification. In *IJCAI*. Citeseer, 1479–1485.

Luigi Malago, Nicolo Cesa-Bianchi, and J Renders. 2014. Online active learning with strong and weak annotators. In *NIPS Workshop on Learning from the Wisdom of Crowds*.

Oscar Reyes, Carlos Morell, and Sebastián Ventura. 2018. Effective active learning strategy for multi-label learning. *Neurocomputing* 273 (2018), 494–508.

Burr Settles. 2009. Active learning literature survey. (2009).

Mohan Singh, Eoin Curran, and Pádraig Cunningham. 2009. Active learning for multi-label image annotation. In *Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science*. 173–182.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems* 159 (2018), 244–258.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009), 2035–2043.

Jian Wu, Victor S Sheng, Jing Zhang, Pengpeng Zhao, and Zhiming Cui. 2014. Multi-label active learning for image classification. In *2014 IEEE international conference on image processing (ICIP)*. IEEE, 5227–5231.

Ming Wu, Qianmu Li, Jing Zhang, Shicheng Cui, Deqiang Li, and Yong Qi. 2017. A robust inference algorithm for crowd sourced categorization. In *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE, 1–6.

Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G Dy. 2011. Active learning from crowds. In *ICML*.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 917–926.

Chen Ye, Jian Wu, Victor S Sheng, Pengpeng Zhao, and Zhiming Cui. 2015. Multi-label active learning with label correlation for image classification. In *2015 IEEE international conference on image processing (ICIP)*. IEEE, 3437–3441.

Guoxian Yu, Jinzheng Tu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. 2020. Active Multilabel Crowd Consensus. *IEEE transactions on neural networks and learning systems* (2020).

Cha Zhang and Tsuhan Chen. 2002. An active learning framework for content-based information retrieval. *IEEE transactions on multimedia* 4, 2 (2002), 260–268.

Jing Zhang, Victor S Sheng, Jian Wu, and Xindong Wu. 2015. Multi-class ground truth inference in crowdsourcing with clustering. *IEEE Transactions on Knowledge and Data Engineering* 28, 4 (2015), 1080–1085.

Jing Zhang and Xindong Wu. 2018. Multi-label inference for crowdsourcing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2738–2747.

Jing Zhang and Xindong Wu. 2019. Multi-Label Truth Inference for Crowdsourcing Using Mixture Models. *IEEE Transactions on Knowledge and Data Engineering* (2019).

Jing Zhang, Xindong Wu, and Victor S Sheng. 2014. Imbalanced multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2014), 489–503.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 7 (2007), 2038–2048.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2013), 1819–1837.

Feipeng Zhao and Yuhong Guo. 2015. Semi-supervised multi-label learning with incomplete labels. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Jinhong Zhong, Ke Tang, and Zhi-Hua Zhou. 2015. Active Learning from Crowds with Unsure Option. In *IJCAI*. 1061–1068.

Zhi-Hua Zhou and Min-Ling Zhang. 2017. Multi-label Learning.