Multi-label Active Learning with Conditional Bernoulli Mixtures

Junyu Chen*, Shiliang Sun*, and Jing Zhao

Department of Computer Science and Technology, East China Normal University,
3663 North Zhongshan Road, Shanghai 200062, P. R. China
juaychen@gmail.com, {slsun,jzhao}@cs.ecnu.edu.cn

Abstract. Multi-label learning is an important machine learning task. In multi-label classification tasks, the label space is larger than the traditional single-label classification, and annotations of multi-label instances are typically more time-consuming or expensive to obtain. Thus, it is necessary to take advantage of active learning to solve this problem. In this paper, we present three active learning methods with the conditional Bernoulli mixture (CBM) model for multi-label classification. The first two methods utilize the least confidence and approximated entropy as the selection criteria to pick the most informative instances, respectively. Particularly, an efficient approximated calculation via dynamic programming is developed to compute the approximated entropy. The third method is based on the cluster information from the CBM, which implicitly takes the advantage of the label correlations. Finally, we demonstrate the effectiveness of the proposed methods through experiments on both synthetic and real-world datasets.

Keywords: Active Learning · Multi-label Classification · Machine Learning.

1 Introduction

Multi-label classification is an important machine learning task and has been used in many aspects of the applications. For many real-world data, one object can be assigned into multiple categories, and the category number of the object is not fixed. This kind of problem is often called multi-label classification. For example, in educational text categorization, the educational news could cover several topics such as preschool, primary school, high school and university. In music information retrieval, a piece of symphony could convey various message such as blue, jazz and classical music. Formally, let $\mathcal{X}$ denote the instance space and $\mathcal{Y} = \{y^1, y^2, \ldots, y^l\}$ denote the label space, the task of multi-label learning is to learn a function $h : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ from the training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, where the power set $\mathcal{P}(\mathcal{Y})$ is the set of all subsets of $\mathcal{Y}$, including the empty set $\emptyset$ and $\mathcal{Y}$ itself. Early multi-label learning mainly focuses on the problem of

* The authors contributed equally to this work.

multi-label text categorization [1–3]. During the past decade, multi-label learning has gradually attracted significant attentions from machine learning and related communities, and has been widely applied to diverse problems such as image automatic annotation [4], web mining [5], tag recommendation [6, 7] etc.

Early researchers on multi-label classification attempt to tackle the task as some well-established learning scenarios. The binary relevance (BR) method decomposes the multi-label learning problem into several independent binary classification problems, where each binary classification problem corresponds to a possible label in the label space [4, 8, 9]. One advantage of the BR method is that the algorithm is easy to implement. The disadvantage is that it ignores the dependence among labels so that the individual label predictions can often be conflicting. For example, in image tagging tasks, an image may be tagged as a cat but not an animal when using the BR method. For dealing with this problem, the power set (PS) method treats each label subset as a class and trains it as a multi-class learning problem [10]. As a consequence, it would be restricted to predicting the label subsets only seen in the training set, and would not predict the labels unseen. Another disadvantage is that the method is often infeasible for the exponential number of labels sets. Recently, the conditional Bernoulli mixtures (CBM) [11] was proposed to be a state-of-the-art multi-label learning method. It is a probabilistic model, which can construct dependencies between labels appropriately.

Given a powerful multi-label learning method, another key point to obtain good performance is to have enough training data or necessary training data. In supervised learning, labeling data is inevitable and tedious. Especially for multi-label learning, the labeling process is much more expensive and time-consuming than single-label problems. Specifically, in the single-label cases, a human annotator only needs to identify a single category to complete labeling, whereas in the multi-label cases, the annotator must consider all the possible labels for each instance, even if the resulting labels are sparse. Thus, if we cannot access the labeled data as many as possible, we can choose the instances to label as necessary as possible. Active learning is to make appropriate instance selection strategies, which aims to choose the most informative instances to obtain the best classification performance. Our work focuses on developing effective active learning methods based on the CBM for solving multi-label learning problems.

There is some existing work on active learning. For example, Gaussian process with manifold-preserving graph reduction (MPGR)-based active leaning (GPMAL) and support vector machine (SVM)-based margin sampling active learning (SVMMAL) are two kinds of active learning methods for binary classification, which provide two kinds of basic guidelines for further research [12]. The GPMAL first applies the MPGR to select a subset and then employs the prediction mean and prediction variance of GP to reselect the most informative instances from the subset. The SVMMAL selects the instances according to the distances from the points to the classification boundary. Besides binary classification active learning methods, some multi-label active learning methods were also developed. Bin-Min [13] was proposed to use the one-versus-all strategy for

multi-label classification with SVM as the base classifier and select the most uncertain instances from the unlabeled set. The mean max loss (MML) or 1DAL strategy [14] selected the instances which had the maximum mean loss value over the predicted classes. In this method, one SVM was trained for each label, and a threshold cutting method was used to decide the target labels. The overall loss value was averaged over the labels. This strategy selected instances only according to the sample correlations, and it did not take advantage of the label correlations. 2DAL [15] considered both relationships between samples and between labels, in which sample-label pairs are chosen to minimize the multi-label Bayesian error bound. More recently, some multi-label active learning methods based on label ranking models were also developed [16–18].

Despite the excellent performance of active learning algorithms, there are still some shortcomings that we should not omit. For example, in the process of active selection, we usually take all the unlabeled instances into account without considering the structural information and spatial diversity among them. This will lead to a result that in the same area there are more than one point to be selected, and thus it is possible to produce redundancy which can decrease the classification accuracy. This phenomenon is also called sampling bias. We will develop effective active learning methods based on CBM with additional sampling bias correction procedure. In addition, In order to avoid the influence of noisy points and simultaneously consider the space connectivity among instances, we introduce a method called cluster-based entropy (CBE) based on CBM. By using CBM, we can construct several clusters which can represent the global distribution structure using fewer instances. This can eliminate the influence of noisy points and promote the selection quality.

In this paper, we propose three principled multi-label active learning methods based on the CBM. The first two multi-label active learning strategies are based on the least confidence and approximated entropy, respectively. The third strategy is made according to the CBE. We evaluate our methods on both the synthetic data and real-world data. The promising experimental results demonstrate the effectiveness of the proposed multi-label active learning methods, and the detailed performance difference among the three proposed methods are also analyzed to give guidance for later researchers.

2 Bernoulli Mixtures and Conditional Bernoulli Mixtures

Bernoulli Mixtures (BM) is a classical model for multi-dimensional binary variable density estimation, where the learnability is realized by assuming independence of variables within each mixture component. Thus, each component density is simply a product of Bernoulli densities, and the overall model has the form

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} Bern(y_l; \mu_{lk}), \tag{1}$$

where μ_{lk} represents the parameters of the lth Bernoulli distribution in the kth component. BM provides an effective approach to model the dependency among different binary variables but with the formulation that is easy to compute.

For multi-label learning, the analysis in [19] depicts that labels could be conditionally independent given input features. With this assumption, conditional Bernoulli mixtures (CBM) [11] extends the BM with both mixture coefficients and Bernoulli distributions conditional on $\boldsymbol{x}$. The distribution of the labels conditional on the input is expressed as

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}; \alpha_k) \prod_{l=1}^{L} Bern(y_l|\boldsymbol{x}; \beta_{kl}), \tag{2}$$

where α_k represents the parameters of function $\pi_k(\cdot)$, and β_{kl} represents the parameters of the lth Bernoulli distribution in the kth component.

The structure of CBM is similar to mixture of experts (ME) [20], where a gate function divides the input space into disjoint regions probabilistically, and an expert model generates the output for their region. We can view CBM as a multi-label extension of mixture of experts with a particular factorization of labels inside each expert. Thus, CBM tackles the multi-label problem as a multi-class problem and several binary problems. The categorical distribution $\pi_k(\boldsymbol{x}; \alpha_k)$ also called gating function assigns each instance $\boldsymbol{x}$ to the kth component with probability $\pi_k(\boldsymbol{x}; \alpha_k)$, which divides the input space into several regions such that each region only contains conditional independent labels. The gating function $\pi_k(\boldsymbol{x}; \alpha_k)$ can be instantiated by any multi-class classifier which provides probabilistic estimate, such as multinomial logistic regression, and the label prediction function $Bern(y^l|\boldsymbol{x})$ can be instantiated by any binary classifier with probabilistic outputs.

In addition, the prediction of CBM is a notable problem, as making the optimal prediction in terms of subset accuracy requires finding the most probable label subset $\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x})$. There are 2^L label subset candidates, and it is intractable to evaluate the probability for each of them. Many multi-label methods suffer from this intractability for exact inference. CBM [11] uses the ancestor sampling strategy for the prediction, where the component index k according to the mixture coefficient $\pi_k(\boldsymbol{x}; \alpha_k)$ is first sampled, and then each label y_l is independently sampled with probability $Bern(y_l|\boldsymbol{x}; \beta_{kl})$. The procedure can be repeated multiple times to generate a set of $\boldsymbol{y}$ candidates, from which we pick the most frequent one. Sampling is easy to implement, but does not guarantee that the predicted $\boldsymbol{y}$ is the global optimal.

3 Methods

3.1 Learning Framework

In order to select queries, an active learner must have a way of assessing how informative each instance is. Let $\boldsymbol{x}^*$ be the most informative instance according to some query strategy $\phi(\boldsymbol{x})$, which is a function used to evaluate each instance

$\boldsymbol{x}$ in the unlabeled pool $\mathcal{U}$. Moreover, let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional instance space and $\mathcal{Y} = \{y^1 \dots y^l\}$ denote the label space, the task of multi-label learning is to learn a function $h : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, where the power set $\mathcal{P}(\mathcal{Y})$ is the set of all subsets of $\mathcal{Y}$ including the empty set and $\mathcal{Y}$ itself.

In many real-world learning problems, lots of unlabeled data are collected at once, and we assume that there is a small set of labeled data $\mathcal{L}$ and a large pool of unlabeled data $\mathcal{U}$. We query a batch of data from unlabeled pool and add them to the labeled set. The overall active learning procedure is described in Algorithm 1.

Algorithm 1 Pool-based active learning.

Input: Labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$, batch size B and informative function $\phi(\boldsymbol{x})$

repeat
  Train classifier $\mathcal{C}$ with labeled set $\mathcal{L}$.
  for $b = 1 \dots B$ **do**
    Query the most informative instance $\boldsymbol{x}_b{}^* = \arg\max_{\boldsymbol{x} \in \mathcal{U}} \phi(\boldsymbol{x})$.
    Labeled set $\mathcal{L} = \mathcal{L} \cup (\boldsymbol{x}_b{}^*, \boldsymbol{y}_b^*)$.
    Unlabeled set $\mathcal{U} = \mathcal{U} \setminus \boldsymbol{x}_b{}^*$.
  end for
until Enough instances are queried

3.2 Selection Criteria

In this section, we will introduce three criteria to select a bunch of informative instances, in which the instance uncertainty or label dependence are considered.

Maximize Least Confidence (LC) For problems with multiple labels, an intuitive selection strategy is to query the instance whose prediction has the least confident:

$$\phi^{LC}(\boldsymbol{x}) = 1 - \arg\max_{\boldsymbol{y} \in \mathcal{P}(\mathcal{Y})} p(\boldsymbol{y}|\boldsymbol{x}). \tag{3}$$

This approach queries the instance for which the current model has the least confidence in its most likely label. However, this criterion only considers information about the most probable label and throws away information about the rest of labels.

Maximize Approximate Entropy (AE) Another uncertainty-based measure of informativeness is entropy [21]. For a discrete random variable X, the entropy is given by

$$H(X) = -\sum_{X} p(X) \ln p(X). \tag{4}$$

In active learning, we wish to employ the entropy of our model's prediction distribution over its labels. Thus, we have

$$\phi^{E}(\boldsymbol{x}) = -\sum_{\boldsymbol{y} \in \mathcal{P}(\mathcal{Y})} p(\boldsymbol{y}|\boldsymbol{x}) \ln p(\boldsymbol{y}|\boldsymbol{x}). \tag{5}$$

However, the number of possible labels grows exponentially with the element number of $\mathcal{Y}$. Empirically, only a few labels contribute to the entropy, and the probabilities for the rest of labels are almost zero which can be ignored. Note that $\lim_{p\to 0} p \ln p = 0$, and we shall take $p(\boldsymbol{y}|\boldsymbol{x}) \ln p(\boldsymbol{y}|\boldsymbol{x}) = 0$. Thus, we have the form of approximate entropy as

$$\phi^{AE}(\boldsymbol{x}) = -\sum\nolimits_{\boldsymbol{y}\in\mathcal{N}(\mathcal{Y})} p(\boldsymbol{y}|\boldsymbol{x}) \ln p(\boldsymbol{y}|\boldsymbol{x}) \leq \phi^{E}(\boldsymbol{x}), \tag{6}$$

where $\mathcal{N}(\mathcal{Y}) = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ is the set of the N most possible labels in the power set. It is worth noting that the number N here is not a fixed number, which indicates a threshold making sure that the sum of probabilities in most possible labels $\sum_{i=1}^{N} p(\boldsymbol{y}_i|\boldsymbol{x})$ is very close to 1. In addition, $\phi^{AE}(\boldsymbol{x})$ is the lower bound of $\phi^{E}(\boldsymbol{x})$, and it will become a tighter bound as $\sum_{i=1}^{N} p(\boldsymbol{y}_i|\boldsymbol{x})$ is closer to 1. Inspired by the dynamic programming prediction method in CBM [11], we present an algorithm for finding the labels with higher probabilities, and calculate the approximate entropy to measure the uncertainty.

To calculate the approximate entropy for $p(\boldsymbol{y}|\boldsymbol{x})$, we need to find the labels with higher probability $p(\boldsymbol{y}|\boldsymbol{x})$. There must exist a component k for which the component probability $\prod_{l=1}^{L} Bern(y_l; \mu_k)$ is high. Thus, we can drop those labels with lower probabilities in each component. We iterate on finding the next label $\boldsymbol{y}$ with the highest probability and add it to the label set until the reset subset candidates will never produce a high probability. The overall procedure is described in Algorithm 2.

Maximize Cluster-Based Entropy (CBE) The two strategies mentioned above only consider the instance uncertainty and do not take advantages of the label correlations. Taking advantages of the simplicity of CBM, we can implicitly capture label correlations rather than directly model such correlations. The labels for the data point $\boldsymbol{x}$ whose mixing coefficients in some components account for a particular proportion often contain correlations. We can verify it by computing the following covariance matrix.

$$\begin{aligned} Cov[\boldsymbol{y}|\boldsymbol{x}] &= \sum_{k=1}^{K} \pi_k [\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k{}^{\top}] - E[\boldsymbol{y}|\boldsymbol{x}] E[\boldsymbol{y}|\boldsymbol{x}]^{\top} \\ &= \sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_k + \sum_{k=1}^{K} \pi_k (1-\pi_k) \boldsymbol{\mu}_k \boldsymbol{\mu}_k{}^{\top} - \sum_{i<j} \pi_i \pi_j \boldsymbol{\mu}_i \boldsymbol{\mu}_j{}^{\top}, \end{aligned} \tag{7}$$

where $E[\boldsymbol{y}|\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k = \mathrm{diag}\{\mu_{lk}(1-\mu_{lk})\}$. Because $\boldsymbol{\Sigma}_k$ is a diagonal matrix, the non-diagonal elements come from the rest part in Equation (7). Those data points only belonging to a single component can be considered as having no dependent labels. Accordingly, we can calculate the entropy of probabilistic gating functions to model such correlations. In order to take account of both least confidence and label correlation and avoid introducing new parameters, we start with the points with highest entropy of $\pi_k(\boldsymbol{x})$ from each clusters, respectively.

Algorithm 2 Approximate Entropy Calculation by Dynamic Programming

Input: Trained CBM model $\mathcal{C}$ and corresponding parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$.

Initialize candidate component set $\mathcal{S} = \{1, 2, \ldots, K\}$, label set $\mathcal{N}(\mathcal{Y})$ and maximum marginal probability M
for $k = 1, 2, \ldots, K$ **do**
  Initialize the maximum component probability G_k.
end for
while $\mathcal{S} \neq \emptyset$ **do**
  for $k \in \mathcal{S}$ **do**
    Find the next highest probability label $\boldsymbol{y}$ unseen in $\mathcal{N}(\mathcal{Y})$ in component k.
    Add label $\boldsymbol{y}$ and corresponding $p(\boldsymbol{y}|\boldsymbol{x})$ to the label set $\mathcal{N}(\mathcal{Y})$.
    Let $p = \sum_{m=1}^{K} \pi_m \prod_{l=1}^{L} Bern(y_l; \mu_m)$ and $q = \prod_{l=1}^{L} Bern(y_l; \mu_k)$.
    if $p > M$ **then**
      Set $M = p$.
    end if
    if $\pi_k q \leq M/K$ or $\pi_k q + \sum_{m \neq k} \pi_m G_m \leq M$ **then**
      Remove k from $\mathcal{S}$.
    end if
  end for
end while
Calculate the approximate entropy $\phi = -\sum_{\boldsymbol{y} \in \mathcal{N}(\mathcal{Y})} p(\boldsymbol{y}|\boldsymbol{x}) \ln p(\boldsymbol{y}|\boldsymbol{x})$

Output: The approximate entropy ϕ.

Then, we reselect the least confident points from each cluster and add those to the labeled set. This method also considers the cluster information and prevents the selected data point far away from the underlying distribution.

3.3 Sampling Bias Correction

In the analysis of [22], the labeled points are always not the representatives of the underlying distribution, because in the setting of active learning, querying the unlabeled point closest to the boundary (or most uncertain, or most likely to decrease overall uncertainty) is very easy to be far away from the underlying distribution due to the presence of noise.

Sampling bias is also one of the most fundamental challenge posed by active learning. LC and AE methods are unable to handle such problem. In this paper, we present a random heuristic method to solve it. Specifically, we equivalently divide the instances in the unlabeled set up into several clusters, and select the most uncertain instance from each cluster and add them to the labeled set. With the benefit of CBM, CBE method inherently contains clustering information and prevents the selected data point far away from the underlying distribution.

4 Experiments

In this section, we present the evaluation on a synthetic dataset and a real-world dataset. We compared the following approaches in our experiments:

- **Random**, the baseline using random selection strategy.
- **LC**, the active learning method based on the strategy of maximizing least confidence.
- **AE**, the active learning method based on the strategy of maximizing approximate entropy.
- **CBE**, the active learning method based on the strategy of maximizing cluster-based entropy.

4.1 Experimental setting

In our experiments, we split the training set and testing set once, randomly select some labeled data point from the training set and let the rest of it become the unlabeled set. We proceed to randomly select the labeled set as the starting training size for ten times and record the average result. CBM is used to train a multi-label classifier for all the comparing active learning methods, in which the gating function $\pi_k(\boldsymbol{z}|\boldsymbol{x};\alpha_k)$ is instantiated by multinomial logistic regression and $Bern(y^l|\boldsymbol{x},\boldsymbol{z};\beta_k)$ is instantiated by logistic regression. We set the number of components to $K = 30$ and the variance of Gaussian prior to $\sigma = 10$. As a baseline, the method of random selection is performed. The real-world dataset used in our experiments is available from the Mulan[1].

We use two metrics to measure the performance of our methods, hamming loss and F1 score. The definitions are as follows.

- Hamming loss:
$$\frac{1}{NL}\sum_{n=1}^{N}\sum_{l=1}^{L}\mathbf{XOR}(\boldsymbol{y}_{nl},\boldsymbol{y}_{nl}{}^{*}),$$
where **XOR** is exclusive (or exclusive disjunction) operation that outputs true only when inputs differ. In practice we substitute true value for one. Hamming loss evaluates the fraction of misclassified instance-label pairs, i.e. a relevant label is missed or an irrelevant is predicted.
- F1 Score:
$$F1 = 2 \times \frac{precision \times recall}{precision + recall}.$$
The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. In our experimental settings, we use micro-F1 metric, which counts the total true positives, false negatives and false positives for each label.

4.2 Datasets and Results

Synthetic Dataset We first consider a simple synthetic dataset with a two-dimensional input $\boldsymbol{x} = (x_1, x_2)$ which are sampled from a mixture of Gaussian distributions with two components. The mean of the two Gaussian distributions

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

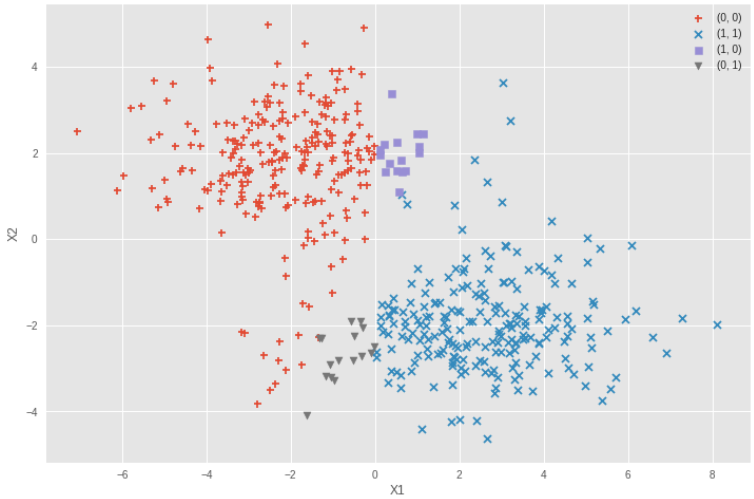

Fig. 1. The distribution of the synthetic 500 instances, where the two labels are rendered with different colors and markers, and the data are linear separable with two linear decision boundaries ($x_1 = 0$ and $x_2 = \sqrt{3}x_1$).

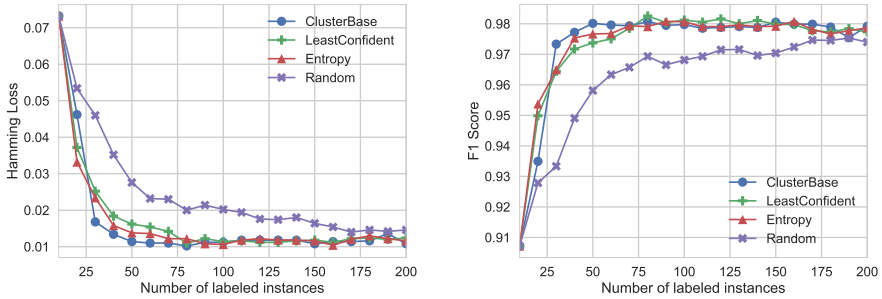

Fig. 2. The average results over 10 runs regarding hamming loss and F1 score on the synthetic dataset.

lies in the second quadrant and the fourth quadrant, respectively. The two labels y_1, y_2 are as follows. The first label is set to one for positive values of x_1, and to zero for negative values, i.e., $y_1 = [x_1 > 0]$. The second label is defined in the same way, but the decision boundary ($x_1 = 0$) is rotated by an angle $\alpha = \pi/3$. The two decision boundaries partition the input space into four regions. It is a two-label classification problem on this dataset. We generate 500 example from the mixture of Gaussian distributions, and we use 250 examples as the training set, 250 examples as the test set. Figure 1 depicts the distribution of the generated data. In the beginning, we randomly select five examples as the labeled set and the rest of examples as the unlabeled set. Figure 2 shows the average results over ten runs regarding F1 score and hamming loss. Table 1 and Table 2 compares the performance on F1 score and hamming loss after selecting 10 unlabeled instances each time, showing the mean and standard deviation

Table 1. Performance in terms of F1 score on the synthetic dataset.

# of selected instances	CBE	AE	LC	RND
20	93.49 ± 2.33	$\mathbf{95.37 \pm 1.83}$	94.99 ± 1.2	92.79 ± 1.56
30	$\mathbf{97.33 \pm 1.77}$	96.49 ± 1.85	96.44 ± 1.8	93.33 ± 1.61
40	$\mathbf{97.72 \pm 0.85}$	97.53 ± 0.94	97.17 ± 1.41	94.91 ± 1.89
50	$\mathbf{98.01 \pm 0.34}$	97.67 ± 0.46	97.37 ± 1.53	95.81 ± 1.64
60	$\mathbf{97.96 \pm 0.18}$	97.68 ± 0.61	97.51 ± 1.09	96.33 ± 1.95
70	$\mathbf{97.93 \pm 0.18}$	97.93 ± 0.82	97.85 ± 0.95	96.57 ± 1.58

Table 2. Performance in terms of hamming loss on the synthetic dataset.

# of selected instances	CBE	AE	LC	RND
20	4.62 ± 1.59	$\mathbf{3.32 \pm 0.94}$	3.72 ± 1.22	5.34 ± 1.68
30	$\mathbf{1.68 \pm 1.09}$	2.34 ± 1.26	2.52 ± 1.25	4.60 ± 1.51
40	$\mathbf{1.34 \pm 0.44}$	1.58 ± 0.62	1.84 ± 0.90	3.52 ± 1.14
50	$\mathbf{1.14 \pm 0.23}$	1.38 ± 0.26	1.62 ± 1.02	2.76 ± 0.99
60	$\mathbf{1.10 \pm 0.17}$	1.36 ± 0.51	1.54 ± 0.78	2.32 ± 1.13
70	$\mathbf{1.10 \pm 0.14}$	1.22 ± 0.61	1.42 ± 0.59	2.30 ± 0.91

of ten-time experiments. The proposed CBE strategy outperforms the other strategies on the whole.

According to Figure 2, we find that CBE obtains slight improvement while the other two methods make a big step at the first selection. It is because that at the beginning the classifier with inadequate data cannot capture the label correlations. CBM overtakes other methods and demonstrates its superiority after the second selection. According to Table 1 and Table 2, it is worthwhile to mention that AE and CBE are more stable than LC, which have lower standard deviation. This is attributed to the integrated consideration of all label combinations in AE and the additional cluster information in CBE.

Scene Dataset We also consider the real-world dataset for evaluation. SCENE is an multi-label image dataset which has 6 labels (beach, sunset, fall foliage, field, mountain, urban). The features are extracted after conversion from raw images to LUV space and are divided into 49 blocks using a 7×7 grid. We compute the first and second moments (mean and variance) of each band, corresponding to a low-resolution image and computationally inexpensive texture features, respectively. The result is a $49 \times 2 \times 3 = 294$-dimensional feature vector per image. We use the default training/test split set for the dataset. In the beginning, we randomly select 80 examples as the labeled set and the rest of examples as the unlabeled set. The experiments are repeated for 10 times and the average results are reported in Figure 3. The classifier achieves high performance with much fewer iterations by our proposed active learning approach. Table 3 and Table 4 compares the performance on F1 score and Hamming loss after selecting 20 unlabeled instances each time. Seen from the over trends of the curve, the proposed cluster-based entropy strategy outperforms the other strategies.

According to the Figure 3, we find that CBE obtains slight improvement at the start again and overtakes other methods after more iterations and CBE is

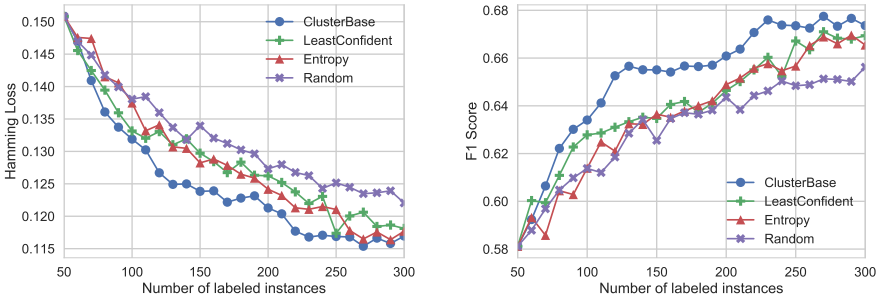

Fig. 3. The average results over 10 runs regarding hamming loss and F1 score on the SCENE dataset.

Table 3. Performance in terms of F1 score on the SCENE dataset.

# of selected instances	CBE	AE	LC	RND
70	**60.64 ± 3.84**	58.58 ± 4.08	59.94 ± 3.14	59.69 ± 4.05
90	**63.02 ± 2.72**	60.28 ± 3.52	62.27 ± 3.76	60.98 ± 2.66
110	**64.12 ± 2.10**	62.48 ± 1.89	62.86 ± 4.24	61.20 ± 1.81
130	**65.66 ± 2.20**	63.25 ± 2.21	63.33 ± 3.24	62.85 ± 1.74
150	**65.51 ± 1.38**	63.65 ± 1.63	63.47 ± 3.24	62.55 ± 2.48
170	**65.67 ± 2.03**	63.79 ± 2.44	64.18 ± 2.35	63.71 ± 2.60
190	**65.70 ± 2.74**	64.21 ± 3.19	64.11 ± 1.22	63.82 ± 2.57
210	**66.37 ± 1.50**	65.16 ± 2.57	65.02 ± 1.52	63.84 ± 2.27
230	**67.60 ± 0.81**	65.77 ± 1.50	66.03 ± 1.76	64.63 ± 2.89
250	**67.36 ± 1.75**	65.67 ± 2.72	66.72 ± 1.39	64.84 ± 2.41
270	**67.75 ± 2.38**	66.89 ± 2.18	67.10 ± 1.09	65.13 ± 1.95
290	**67.66 ± 2.36**	66.96 ± 1.53	66.80 ± 1.28	65.02 ± 2.93

the most efficient method among them. Being different from the one in synthetic dataset, the report in scene dataset shows that LC is more effective than AE and encounter some difficulties at the beginning. It is because the entropy can be inaccurate when the label space is very large and the classifier is not well developed. In general, the CBE is the most effective and stable method among the three methods.

4.3 Discussion

From the above two experiments, we observe that the proposed CBE strategy has a slight improvement at the beginning compared to other two methods but has a more significant improvement after several iterations. This phenomenon can be attributed to the fact that at the beginning the classifier with inadequate data cannot capture the label correlations. After several iterations, the classifier has constructed label dependence to a certain extent and is more likely to select such informative instances that contain more than one label. Therefore, it is reasonable to employ the least confidence or the entropy strategy at the start and shift to the cluster-based entropy strategy when the classifier has captured label dependence. The combining method may boost the performance, and the

Table 4. Performance in terms of hamming loss on the SCENE dataset.

# of selected instances	CBE	AE	LC	RND
70	**14.09 ± 1.19**	14.74 ± 1.31	14.25 ± 1.03	14.48 ± 1.41
90	**13.37 ± 0.80**	14.06 ± 1.26	13.60 ± 1.19	13.99 ± 0.94
110	**13.02 ± 0.84**	13.32 ± 0.69	13.20 ± 1.37	13.84 ± 0.74
130	**12.49 ± 0.76**	13.07 ± 0.68	13.10 ± 1.02	13.37 ± 0.69
150	**12.38 ± 0.35**	12.82 ± 0.55	12.97 ± 1.01	13.39 ± 0.90
170	**12.22 ± 0.57**	12.78 ± 0.77	12.67 ± 0.76	13.12 ± 0.65
190	**12.31 ± 0.76**	12.59 ± 1.10	12.63 ± 0.44	12.96 ± 0.66
210	**12.04 ± 0.36**	12.32 ± 0.81	12.52 ± 0.60	12.80 ± 0.66
230	**11.68 ± 0.30**	12.11 ± 0.47	12.20 ± 0.61	12.63 ± 0.85
250	**11.69 ± 0.53**	12.10 ± 0.97	11.74 ± 0.49	12.52 ± 0.66
270	**11.54 ± 0.77**	11.65 ± 0.73	12.06 ± 0.54	12.35 ± 0.57
290	**11.58 ± 0.75**	11.64 ± 0.55	11.86 ± 0.63	12.39 ± 0.88

conjecture also need more experiments to verify. Besides, it also needs a strategy to find the appropriate time to shift. Finally, for most cases, we observe that CBE is comparable to AE and LC, which means that the performance of CBE is more stable.

5 Conclusion and Future Work

In this paper, we present three active learning algorithms for multi-label classification with CBM. It utilizes least confidence, approximated entropy and cluster-based entropy as the uncertainty measure to pick the most informative data points. Experimental results on synthetic and real-world datasets show that our methods outperform random selection. In the future, we plan to perform theoretical analysis on the methods and extend our work to multiview multi-label active learning [23–25].

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Project 61673179 and Shanghai Sailing Program. The corresponding author is Jing Zhao.

References

1. McCallum, A.: Multi-label text classification with a mixture model trained by EM. In: AAAI workshop on Text Learning. pp. 1–7 (1999)
2. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning **39**(2-3), 135–168 (2000)
3. Ueda, N., Saito, K.: Parametric mixture models for multi-labeled text. In: Advances in Neural Information Processing Systems. pp. 737–744 (2003)
4. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9), 1757–1771 (2004)
5. Kazawa, H., Izumitani, T., Taira, H., Maeda, E.: Maximal margin labeling for multi-topic text categorization. In: Advances in Neural Information Processing Systems. pp. 649–656 (2005)

6. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: Proceedings of the ECML/PKDD. pp. 1–9 (2008)
7. Song, Y., Zhang, L., Giles, C.L.: A sparse Gaussian processes classification framework for fast tag suggestions. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. pp. 93–102 (2008)
8. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning. pp. 137–142 (1998)
9. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5**(4), 361–397 (2004)
10. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing & Mining **3**(3), 1–13 (2007)
11. Li, C., Wang, B., Pavlu, V., Aslam, J.: Conditional Bernoulli mixtures for multi-label classification. In: Proceedings of the International Conference on Machine Learning. pp. 2482–2491 (2016)
12. Zhou, J., Sun, S.: Gaussian process versus margin sampling active learning. Neurocomputing **167**, 122–131 (2015)
13. Brinker, K.: On active learning in multi-label classification. From Data and Information Analysis to Knowledge Engineering pp. 206–213 (2006)
14. Li, X., Wang, L., Sung, E.: Multilabel SVM active learning for image classification. In: Proceedings of the International Conference on Image Processing. pp. 2207–2210 (2004)
15. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Zhang, H.J.: Two-dimensional active learning for image classification. In: Computer Vision and Pattern Recognition. pp. 1–8 (2008)
16. Huang, S., Zhou, Z.: Active query driven by uncertainty and diversity for incremental multi-label learning. In: IEEE 13th International Conference on Data Mining. pp. 1079–1084 (2013)
17. Huang, S., Chen, S., Zhou, Z.: Multi-label active learning: query type matters. In: International Joint Conference on Artificial Intelligence. pp. 946–952 (2015)
18. Gao, N., Huang, S., Chen, S.: Multi-label active learning by model guided distribution matching. Frontiers of Computer Science **10**(5), 845–855 (2016)
19. Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: On label dependence and loss minimization in multi-label classification. Machine Learning **88**(1-2), 5–45 (2012)
20. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Neural Computation **6**(2), 181–214 (1994)
21. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948)
22. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: Proceedings of the International Conference on Machine Learning. pp. 208–215 (2008)
23. Sun, S.: A survey of multi-view machine learning. Neural Computing and Applications **23**, 2031–2038 (2013)
24. Zhao, J., Xie, X., Xu, X., Sun, S.: Multi-view learning overview: Recent progress and new challenges. Information Fusion **38**, 43–54 (2017)
25. Sun, S., Shawe-Taylor, J., Mao, L.: PAC-Bayes analysis of multi-view learning. Information Fusion **35**, 117–131 (2017)