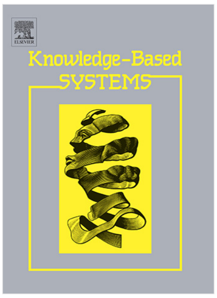
An efficient active learning method for multi-task learning

Yanshan Xiao, Zheng Chang, Bo Liu



PII: S0950-7051(19)30495-2
DOI: https://doi.org/10.1016/j.knosys.2019.105137
Reference: KNOSYS 105137

To appear in: *Knowledge-Based Systems*

Received date : 29 January 2019
Revised date : 17 October 2019
Accepted date : 19 October 2019

An Efficient Active Learning Method for Multi-task Learning

Yanshan Xiao[a], Zheng Chang[a,*], Bo Liu[b]

[a] Faculty of Computer, Guangdong University of Technology, China
[b] Faculty of Automation, Guangdong University of Technology, China

Abstract

In multi-task learning, the sharing of information between related tasks affects and promotes each other's learning. However, the traditional multi-task learning techniques are based on sufficient labeled data to improve the learning of each task, and labeling samples is always expensive in practice. In this paper, we propose two variants active learning methods for multi-task classification problems. In the uncertainty step, we propose the support vector preservation criterion that evaluates uncertainty at the level of classifier, called classifier-level uncertainty (CLU). In the diversity step, we propose two diversity criteria that evaluate diversity by the clustering method and the partition method respectively, called clustering-based diversity (CBD) and partition-based diversity (PBD). Each diversity criterion is together with the uncertainty criterion to form an active learning method for multi-task learning. In addition, the proposed support vector preservation criterion selects local informative samples which determine the hyperplane for each task. Furthermore, in order to maintain the distribution structure of the samples, we put forward the micro-kernel k-means clustering method and partition-based method to select global informative samples from the non-support vectors. By incorporating the local and global informative samples into active learning, we propose two active learning methods for multi-task problems. We evaluate the effectiveness of the proposed methods by conducting experiments with other active learning methods. The experimental results show that the proposed two methods perform better than other active learning methods.

Keywords: Multi-task classification, Active learning, Support vector machine

[*] Corresponding author

Email addresses: xiaoyanshan@189.cn (Yanshan Xiao), 854494092qq@sina.com (Zheng Chang), csboliu@163.com (Bo Liu)

1. Introduction

Traditional learning focuses on a single task, which causes other information that optimizes the model performance to be ignored. Multi-task learning (MTL) is different from the traditional learning. By sharing the characterization between the related tasks, the model can be better generalized. Since the shared representation and multiple tasks are performed simultaneously, the number of the training samples and the overall model parameters are reduced, which makes the task execution more efficient. In recent years, support vector machines[1, 2] (SVM) have been successfully used in multi-task classification. The SVM-based methods are designed based on a general assumption that the classifier model parameter values of different tasks are close to each other[3]. Sharing parameters and representations in SVM improve the performance of each classifier. Multi-task learning is widely used in various fields, such as drug-drug interaction extraction[4], person identification[5–7] and object tracking[8, 9], etc.

The previous work on multi-task learning can be classified into several categories based on learning algorithms, including SVM-based multi-task learning[10–12], decision tree-based multi-task learning[13, 14], neural networks-based multi-task learning[15–18], and Bayesian multi-task learning[19–21], etc. Specifically, in SVM-based multi-task learning, they aim to get classifiers by solving a quadratic programming problem. For example, the multi-task learning formulations based on support vector machines are presented in [10] on the assumption that models or model parameters of tasks are close to a certain mean function. In decision tree-based multi-task learning, a multi-task decision tree is used to make a final decision for some tasks in an internal test node, when the internal test node has enough information to classify an instance of a certain task. For example, a novel technique for the multi-task learning called Multi-Task Decision Tree (MT-DT) is proposed in [13]. This technique can benefit from an improved information gain criterion. In neural networks-based multi-task learning, the generalization ability of the model is improved through soft sharing and hard sharing of hidden layer parameters. For example, Huang et al.[15] propose a new multi-task deep neural networks model (MTL-DNN) to denoise for a robust ASR system. In Bayesian multi-task learning, Bayesian methods are used to learn latent structures hiding in the data and infer the task and feature correlations simply. In [19], a generic framework of Bayesian max-margin MTL is proposed and extended to learn latent feature representations, in which structured priors are used to jointly estimate task correlations and feature correlations.

Traditional supervised learning methods are dependent on adequate labeled samples to improve the generalization ability of classification models. However, in real world, the number of labeled samples is always limited and it is time consuming and expensive to get large amounts of labeled samples. Active learning is one of the most effective methods to solve such problems. Different from traditional supervised learning methods, active learning[22, 23] is an iterative process. The goal of active learning is to obtain a well learning model with a small number of labeled samples. During each iteration, an initial classification model is created on a small number of labeled samples. Then, the most informative samples are selected from the unlabeled pool based on a selection strategy. Newly labeled samples are labeled and added to the labeled set, which is used to create a new classification model. In this way, active learning can obtain high classification accuracy with few labeled samples, which reduces the cost of obtaining labeled data. In the past few years, active learning methods are exploited in many research areas. For example, for multi-label classification, Yang et al.[24] propose a novel multi-label active learning approach. In this approach, the unlabeled samples which can lead to the largest reduction of the expected model loss are selected as queries. For supervised classification, CSAL[25] is proposed to define the reliable training sets for the classification of remote sensing images with support vector machines. For multi-class classification, MC_SVMA[26] is proposed to allow active learning to participate in the initial pattern classes mining and the subsequent SVM training.

Although there has been a lot of research on the SVM-based multi-task learning, little research has been done to the active learning techniques for multi-task learning. In order to overcome the problem of SVM-based multi-task learning on the limited labeled samples, we propose two novel active learning methods for the SVM-based multi-task classification. Both active learning methods consist of two steps: the uncertainty step and the diversity step. In the first step, we put forward to a support vectors preservation criterion to select the most informative samples around the classifier for each task. In the second step, we propose two variants of diverse samples selection methods to select the samples which can describe the structure of the data distribution. In all, the main contributions of the paper are as following:

1. In the uncertainty step, we propose the support vectors preservation method to select the samples which have most effect on the classifier for each task. In addition, this is called classifier-level uncertainty criterion (CLU), which evaluates uncertainty at the level of SVM classifier. In this step, we first analyze the objective model of SVM-based multi-task learning, and then select the support vectors which can form the classifier for each task. To preserve the structure of the data distribution, another part of the most uncertain samples are selected in the diversity step.
2. In the diversity step, we propose two different diverse samples selection criteria: 1) clustering-based diversity (CBD) and 2) partition-based diversity (PBD). In CBD, the non-support vectors are divided into different clusters based on the micro-kernel k-means clustering method for each task. For each micro-cluster, we select only one representative sample to form the new training set, which can vaguely maintain the structure of the samples except for support vectors. In PBD, we first divide

the non-support vectors into different partitions in the feature space based on their distances to the hyperplane. We then select one representative sample from each non-empty partition to form a new training set, which can be used in the subsequent active learning. In addition, both diverse samples selection methods can be together with the uncertainty criterion to form a novel multi-task active learning method. In this way, we propose two variants of active learning methods, which are called CLU-CBD and CLU-PBD, for SVM-based multi-task learning.

3. In order to evaluate the effectiveness of the proposed methods, we perform extensive experiments on multiple data sets. In addition, we verify the proposed methods' efficiency by comparing their experimental results to other active learning methods empirically. By comparison, we observe that the proposed methods result in better accuracy with respect to other methods.

The rest of the paper is organized as follows. Section 2 discusses the related work. The active learning strategies and the multi-task SVM are presented in detail in Section 3. Section 4 describes the experimental setup and the experimental results. The conclusion and future work are presented in Section 5.

2. Related Work

In this section, we briefly review previous work related to our research. In section 2.1, we review the previous work on active learning. Then, we review the previous work on multi-task learning in section 2.2.

2.1. Active learning

In the field of machine learning, high-quality samples can make the training process more efficient and less expensive. Active learning[27–29] is an effective learning method that enables high-quality samples to be selected during the training process for reducing sample redundancy and improving the performance of classification model with a small number of labeled samples. Active learning is an iterative process[30], in which the most informative samples are selected from the unlabeled pool, labeled and added to the training set during each iteration. Because the selection strategy directly affects learning efficiency and generalization performance, designing the selection strategies is the crucial step for active learning. According to the query strategies used, active learning can be divided into three categories: uncertainty sampling methods[31–34] and query by committee methods[35–37]. Next, we review the previous work of active learning.

For uncertainty sampling-based strategies, they focus on selecting the samples that have most influence on the current classifier, i.e. the sample which the classifier has low confience in[38], or the samples which are easily mispredicted by the classifier[39], etc.

Culotta and McCallum[31] propose a new active learning paradigm in which each unlabeled instance first is ranked by its confidence value given by the current extractor, then the instance with the least confidence is selected to be labeled by current extractor, and the user is allowed to correct the prediction with true labeling. Balcan et al.[34] present and analyze a generic margin-based active learning framework for learning linear separators and instantiate it for a few important cases. In this framework, weight vectors are learned from the labeled samples and used to label other unlabeled samples that meet the linear condition.

For query by committee-based strategies, the Query by Committee algorithm[40] (QBC) uses the examples, whose expected information gain is high, as queries examples. And it filters the informative examples from the random unlabeled examples that it gets from the oracle, rather than constructing the examples. In [35], it is proposed that the QBC is an efficient query algorithm for the perceptron concept class with distributions that are close to uniform. In addition, when the number of queries increases, the prediction error is guaranteed to decrease if the queries have high expected information gain. Majidi and Crane[36] provide a novel framework in which a committee of parsers are used to generate separate models. These separate models predict the head nodes and the relation to the heads of instances in the unlabeled set. Then, for each sentence with most entropy value, the tokens with the highest entropy are annotated by the expert.

In addition, some methods focus on the impact of current sample selection on future model performance. Thus, the training samples, which are expected to result in the lowest error on future test examples, are selected to be labeled and added to the training data[41]. For example, the method in [42] takes advantage of the fact that an unbiased learner, which minimizes the expected error given as the expected sum of squared error, is equivalent to an unbiased learner, which minimizes its variance. In this method, samples that maximize the error reduction by minimizing the learner's variance are selected as queries. Similarly, the method[43] estimates future error rate either by the entropy of the posterior class distribution on an unlabeled sample, or by the posterior probability of the most probable class for the unlabeled samples.

2.2. Multi-task learning

The traditional single task learning methods only focus on the information of the learning task itself and pay little attention to the information of the related tasks. In fact, by learning related tasks simultaneously, the objective of getting better generalization accuracy can be achieved[44, 45]. Therefore, multi-task learning[46, 47] is studied to address problems of multiple tasks by sharing useful information (such as a common representation or some model parameters) between related tasks. According to the learning algorithms used in multi-task learning, multi-task learning can be divided into three categories:

SVM-based multi-task learning[48–50], neural networks-based multi-task learning[51–53], and Bayesian multi-task learning[54–56].

For SVM-based multi-task learning, support vector machine is used as the classification model and the multi-task problem is transformed into a quadratic optimization problem by parameters sharing and common feature representation[57]. Thus, performance of the classification model can be improved by jointly performing multiple learning tasks. Zhu et al.[48] develop the infinite latent support vector machines (iLSVM) and multi-task infinite latent support vector machines (MT-iLSVM), which combine the discriminative large-margin idea with a nonparametric Bayesian model to learn infinite latent feature models[58] for classification and multi-task learning, respectively. In [49], Jebara first computes a common feature selection or kernel selection configuration for multiple support vector machines trained on different yet inter-related datasets. Then, a multi-task representation learning approach is derived using the maximum entropy discrimination formalism.

For neural networks-based multi-task learning, the algorithms are usually designed to reduce the risk of over-fitting or achieve similarity of parameters by different parameters sharing mechanisms[59]. Liu et al.[52] introduce a gating mechanism, which can better control the information passed by the neuron in shared layers. In addition, they integrate RNN into the multi-learning framework for text classification to map arbitrary text into semantic vector representations. They also propose three models with different shared mechanisms, called uniform-layer architecture, coupled-layer architecture, and shared-layer architecture. Zhang et al.[53] apply MTL combined with CNN to facial landmark detection. And they propose early stopping for multi-task learning to solve the optimization problems caused by different convergence speeds of each task.

For Bayesian multi-task learning, Bayesian approaches are adopted to recognize parallel tasks and learn their underlying structure from the data. In addition, some model parameters are shared and others are soft-shared through a prior distribution in some Bayesian approaches in multi-task learning settings[56, 60]. In [54], Lazaric and Ghavamzadeh first adopt the Gaussian process temporal-difference value function model. Then, they use a hierarchical Bayesian approach to model the distribution over the value functions in the case where the tasks share structure in their value functions. And two hierarchical Bayesian models[54] are presented for two different cases where all the value functions belong to the same class and where they belong to an undefined number of classes. In [55], a novel MTRL approach is proposed, in which Bayesian reinforcement learning is extended to a multi-task setting based on the Bayesian framework. In this approach, a posterior probability distribution is maintained over possible Markov Decision Processes[61] (MDPs). The hierarchical model parameters are initialized to uninformed values. Then, a set of MDPs are generated according to the Dirichlet Process[62] priors and the one with the highest probability is returned to update the model.

Since most of the previous work has focused on a single domain and little work has been done on active learning for multi-task learning. In this paper, two novel active learning methods for multi-task learning are proposed based on both uncertainty criterion and diversity criterion to cope with multi-task classification problems in the absence of sufficient labeled samples. In addition, the uncertain samples are selected at the uncertainty step and the diversity step respectively.

3. The Proposed Active Learning Methods for SVM-based Multi-task Learning

In Section 3.1, the multi-task support vector machine is first presented. In Section 3.2, the basic idea of the proposed active learning methods is presented.

3.1. Multi-task SVM classification

We denote the data set in the kth task as $X_k = L_k \cup U_k$, where L_k and U_k represent the labeled data set and the unlabeled data set in the kth task, respectively. Besides, we have all these tasks on the same space χ, with $\chi \subseteq \mathbb{R}^d$. The goal is to learn n decision functions $f_1(x), f_2(x), ..., f_n(x)$, one for each task.

The ith sample in the kth task is denoted as x_{ik}, the decision function is $f_k(x_{ik}) = w_k \cdot \phi(x_{ik}) + b_k$, in which $\phi(x)$ is a non-linear feature mapping and b_k is the offset vector. if $f_k(x_{ik}) \geq 0, x_{ik}$ is labeled as positive; otherwise, it is labeled as negative. w_k is the normal vector to the decision hyperplane and consists of two parts. The first part is the common mean vector shared by each task, which is denoted as w_0, and the second part is the specific vector v_k for a specific task. Here, w_k for each task is expressed as:

$$w_k = w_0 + v_k, \tag{1}$$

Following the above assumption, SVM can be generalized to multi-task learning. The primal optimization problem can be written as follows:

$$\begin{aligned} \min_{w_0, b_k, v_k, \xi_{ik}} & \frac{1}{2}\|w_0\|^2 + \sum_{k=1}^{n} \lambda_k \|v_k\|^2 + \mathcal{C} \sum_{k=1}^{n} \sum_{i=1}^{n_k} \xi_{ik} \\ \textit{s.t.} \quad & y_{ik}(w_k^{\mathrm{T}} \cdot \phi(x_{ik}) + b_k) \geq 1 - \xi_{ik}, \xi_{ik} \geq 0 \end{aligned} \tag{2}$$

where n_k is the number of data in kth task, $\mathcal{C}$ is penalty parameter that balances the margin and errors. In addition, parameter λ_k is used to control the preference of the tasks. In other words, the larger the value of the parameter λ_k, the higher the preference of the kth task. ξ_{ik} is the corresponding slack variable.

By introducing Lagrangian multipliers α_{ik} for the samples in each task, the solution of Problem (2) is to resolve the dual problem:

$$\begin{aligned} \max_{\alpha_{ik}} & \sum_{k=1}^{n}\sum_{i=1}^{n_k}\alpha_{ik} - \frac{1}{2}\sum_{k=1}^{n}\sum_{t=1}^{n}\sum_{i=1}^{n_k}\sum_{j=1}^{n_t}\alpha_{ik}\alpha_{jt}\langle\phi(x_{ik}),\phi(x_{jt})\rangle \\ \textit{s.t.} \quad & 0 \le \alpha_{ik} \le \mathcal{C} \end{aligned} \tag{3}$$

Suppose that we define a kernel function as $k(x_{ik}, x_{jt}) = \langle\phi(x_{ik}), \phi(x_{jt})\rangle$, where k and t are the task index associated to each sample. Thus, the classifier of each task is obtained by solving the above optimization problem using the kernel function $k(x_{ik}, x_{jt})$, and then the decision function for each task is given by:

$$f_k(x) = \sum_{k=1}^{n}\sum_{i=1}^{n_k}\alpha_{ik}k(x_{ik}, x) + b_k \tag{4}$$

3.2. The Proposed Methods

In this section, the proposed active learning strategies are presented by dividing the strategy into two consecutive steps: the uncertainty step and the diversity step. In the uncertainty step, the uncertain samples around the hyperplane are extracted. In the diversity step, the uncertain sample outside the margins are selected.

3.2.1. Uncertainty step

The uncertainty criterion identifies samples with the lowest classification confidence as the most uncertain samples, since the samples, which the classifier is most uncertain with, always have useful information for the classifier. We propose a novel uncertainty criterion, called support vectors preservation criterion, which selects the samples around the classifier for each task in the SVM-based multi-task learning. Since the proposed uncertainty criterion evaluates uncertainty at the level of SVM classifier, it can be abbreviated as CLU. Details are as follows.

Take the case of two-task, that is transfer learning, as an example, we present the uncertainty criterion for two-task learning. Figure 1 shows an intuitive example in which the uncertain samples are selected based on the uncertain criterion. For the source task in the left of Figure 1, circles and squares represent positive and negative samples, respectively. Shapes filled with blue represent the selected samples for the source task. The solid blue line represents the classification hyperplane and the equation $w_1 \cdot \phi(x) + b_1 = 0$ represents the classifier for the source task. The dotted blue lines represent the margin boundaries, $w_1 \cdot \phi(x) + b_1 = 1$ and $w_1 \cdot \phi(x) + b_1 = -1$. For the target task in the right of Figure 1, triangles and diamonds represent positive and negative samples, respectively. Shapes filled

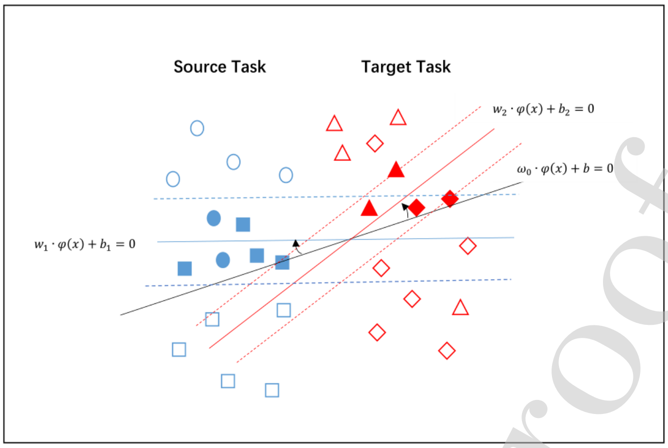

Source Task
Target Task
$w_2 \cdot \varphi(x) + b_2 = 0$
$\omega_0 \cdot \varphi(x) + b = 0$
$w_1 \cdot \varphi(x) + b_1 = 0$

Figure 1: Extraction of the uncertain samples. Color-filled samples are support vectors which determine the classifiers.

with red represent the selected samples for the target task. The solid red line represents the classification hyperplane and the equation $w_2 \cdot \phi(x) + b_2 = 0$ represents the classifier for the target task. The dotted red lines represent the margin boundaries, $w_2 \cdot \phi(x) + b_2 = 1$ and $w_2 \cdot \phi(x) + b_2 = -1$. It can be seen that the solid red and blue samples support the transfer learning classifiers.

For SVM-based multi-task learning, we have the analysis as follows. According to the decision function (4), the samples, which have non-zero α_{ik}, determine the classification hyperplane. Such samples are known as support vectors (SV). Other samples whose Lagrangian coefficient α_{ik} is zero, do not affect the classification hyperplane. The Karush-Kuhn-Tucker Conditions of transfer learning can be written as:

$$\alpha_{ik} = 0 \Leftrightarrow y_{ik}(w_k \cdot \phi(x_{ik}) + b_k) \geq 1 \tag{5}$$

$$0 < \alpha_{ik} < \mathcal{C} \Leftrightarrow y_{ik}(w_k \cdot \phi(x_{ik}) + b_k) = 1 \tag{6}$$

$$\alpha_{ik} = \mathcal{C} \Leftrightarrow y_{ik}(w_k \cdot \phi(x_{ik}) + b_k) \leq 1 \tag{7}$$

Furthermore, for the formula (5), the equation $y_{ik}(w_k \cdot \phi(x_{ik}) + b_k) \geq 1$ means the sample x_{ik} lies outside the margin boundaries. The α_{ik} of samples that lie outside the margin boundaries are zero. For the formula (6)(7), the equation $y_{ik}(w_k \cdot \phi(x_{ik}) + b_k) = 1$ means the sample x_{ik} resides on the margin boundaries, the equation $y_{ik}(w_k \cdot \phi(x_{ik}) + b_k) \leq 1$ means the sample x_{ik} resides inside the margin boundaries. Thus, the samples that lie on or inside the margin boundaries have non-zero α_{ik}. Based on the above analysis, we draw a conclusion that the support vectors (SVs), which fall within the margin boundaries of the source task, determine the classifier of the source task and are easily misclassified by the classifier. The analysis is similar for the target task, the support vectors (SVs), which fall within the margin boundaries of the target task, determine the classifier of the target task and are easily misclassified by the classifier. Adding such samples into the training set to retrain the classifiers will improve the performance of both classifiers.

For all $x_{ik} \in U_k$, the uncertain samples, which are also called support vectors (SVs) here, are selected as

$$SV_k = \{x_{ik} | f_k(x_{ik}) \in [-1, +1]\} \tag{8}$$

where SV_k is the set of samples selected from the unlabeled set of kth task in the uncertainty step. Then, the labels are assigned to the selected samples, the unlabeled set is updated and the selected samples are added to the training set.

3.2.2. Diversity step

In order to vaguely preserve the sample distribution and avoid degradation of the generalization performance in classifying test unlabeled samples, it is necessary to select a part of samples outside the margin boundaries to maintain the basic structure of the data. Next, we propose two different diversity criteria to select the samples outside the margin boundaries in the following diversity step.

A. Clustering-based diversity

Clustering techniques evaluate the distribution of the samples in a feature space and group the similar samples into the same clusters. Since the samples within the same cluster are correlated and provide similar information, selecting a representative sample from each cluster prevents sample redundancy. Here we conduct the micro-kernel k-means clustering algorithm to divide the unlabeled samples outside the margin boundaries into different clusters. Specific steps are as follows:

- For each task, the unlabeled samples U_k are divided into two parts: a positive sample set U_k^+ and a negative sample set U_k^- $(k = 1, 2, ..., n)$, which are denoted as

$$U_k^+ = \{x_{ik} \mid f_k(x_{ik}) > 1, x_{ik} \in U_k\} \tag{9}$$
$$U_k^- = \{x_{ik} \mid f_k(x_{ik}) < -1, x_{ik} \in U_k\} \tag{10}$$

- Apply the micro-kernel k-means clustering algorithm to U_k^+ and U_k^- with $K = h$, respectively. Then, U_k^+ is divided into $h = \frac{|U_k|}{a}$ different clusters $C_1^+, C_2^+, ..., C_h^+$. Similarly, U_k^- is divided into h different clusters $C_1^-, C_2^-, ..., C_h^-$.
- After $2h$ clusters of each task are obtained, we choose the most uncertain sample, which has minimum $|f(x)|$ value, as the representative sample from each cluster.

Then, the $2h$ uncertain samples are selected as

$$x_{tk}^{+} = \arg \min_{x_{ik} \epsilon C_t^{+}} |f(x_{ik})| \tag{11}$$

$$x_{tk}^{-} = \arg \min_{x_{ik} \epsilon C_t^{-}} |f(x_{ik})| \tag{12}$$

where $t = 1, 2, ..., h$, x_{tk}^{+} is the uncertain sample selected from cluster C_t^{+} of kth task, x_{tk}^{-} is the uncertain sample selected from cluster C_t^{-} of kth task.

- After obtaining $2h$ selected samples from the unlabeled samples outside the margin boundaries in each task, assign the labels to the selected samples and add them to the training set. At last, the classifiers get retrained with the updated training set.

B. Partition-based diversity

In order to select unlabeled samples diverse and preserve the basic data structure of the data, we propose a partition-based diversity method that divides the unlabeled samples outside the margin boundaries into different partitions in the feature space. Firstly, the unlabeled set updated in the uncertainty step for each task, such as kth task, is divided into two subsets U_k^{+} and U_k^{-} using (9) and (10). The maximum decision value $f_k^{max}(\cdot)$ of samples in U_k^{+} and the minimum decision value $f_k^{min}(\cdot)$ of samples in U_k^{-} are obtained. Then, each of the two subsets is divided into $m = \frac{|U_k|}{a}$ partitions, the width of each partition in the subsets U_k^{+} and U_k^{-} is computed as

$$W_k^{+} = \frac{f_k^{max} - 1}{m} \tag{13}$$

$$W_k^{-} = \frac{-1 - f_k^{min}}{m} \tag{14}$$

For W_k^{+}, $f_k^{max} - 1$ means the distance between the hyperplane $w_k \cdot \phi(x) + b_k = 1$ and the sample with the largest decision value. For W_k^{-}, $-1 - f_k^{min}$ means the distance between the hyperplane $w_k \cdot \phi(x) + b_k = -1$ and the sample with the smallest decision value. Thus, the lower bounds and upper bounds of each partition can be denoted below:

$$L_k^{+}(i) = 1 + (i-1)W_k^{+} \quad and \quad H_k^{+}(i) = 1 + (i)W_k^{+} \tag{15}$$

$$L_k^{-}(i) = -1 + (i-1)W_k^{-} \quad and \quad H_k^{-}(i) = -1 + (i)W_k^{-} \tag{16}$$

where $i = 1, 2, ..., m$, $L_k^{+}(i)$ and $H_k^{+}(i)$ represent the lower and upper bounds of ith partition of subset U_k^{+}, $L_k^{-}(i)$ and $H_k^{-}(i)$ represent the lower and upper bounds of ith partition of subset U_k^{-}. Let P_k be the set of non-empty partitions of two subsets U_k^{+} and

U_k^-, $P_k(j)$ is jth partition in P_k. Then, we select one sample as representative sample from each non-empty partition as follows:

$$x_{jk} = \arg\min_{x_{ik} \in P_k(j)} |f_k(x_{ik})|, j = 1, 2, ..., |P_k| \tag{17}$$

After obtaining a batch of uncertain samples from the unlabeled samples outside the margin boundaries in each task, they are labeled and added to the training set. Then, the classification model is retrained with the updated training set.

Each diversity technique can be together with the uncertainty technique to form an active learning method for SVM-based multi-task learning. The two active learning methods are called as: CLU with CBD (denoted by CLU-CBD) and CLU with PBD (denoted by CLU-PBD). The details of CLU-CBD and CLU-PBD algorithms for multi-task classification are presented in Algorithms 1 and 2 respectively.

Algorithm 1 CLU-CBD

```
Input: Labeled set $L_k$, unlabeled set $U_k$ for multi-task learning, parameter $a$, the number of tasks $n$.
Output: Classifier for each task $f_0, f_1, ..., f_n$.
Train the classifiers $f_0, f_1, ..., f_n$ for each task on the initial labeled set based on the optimization (2).
repeat
    for $k = 1$ to $n$ do
        Select the set of support vectors using (8) and assign labels to them.
        Add the selected samples into the training set and update the unlabeled set.
        Divide the unlabeled set into two sets $U_k^+$ and $U_k^-$ by (9) and (10).
        Apply micro-kernel k-means clustering algorithm to $U_k^+$ and $U_k^-$ to divide them into $h = \frac{|U_k|}{a}$ different clusters, respectively.
        Choose the representative sample from each non-empty partition in $P_k$ by (17) and assign labels to them.
        Add the selected samples into the training set and update the unlabeled set.
        Retrain the classifiers with the updated training set using optimization (2).
    end for
until stop criterion is satisfied
```

In each iteration of the algorithm CLU-CBD, we first create the initial classifiers with the labeled samples, then we select the uncertain samples inside the margin boundaries using CLU and select the uncertain samples outside the margin boundaries using CBD. The selected uncertain samples are labeled and added into the labeled set. The classifiers

Algorithm 2 CLU-PBD

```
Input: Labeled set $L_k$, unlabeled set $U_k$ for multi-task learning, parameter $a$, the number of tasks $n$.
Output: Classifier for each task $f_0, f_1, ..., f_n$.
Train the classifiers $f_0, f_1, ..., f_n$ for each task on the initial labeled set using the optimization (2).
repeat
    for $k = 1$ to $n$ do
        Select the set of support vectors using (8) and assign labels to them.
        Add the selected samples into the training set and update the unlabeled set.
        Divide the unlabeled set into two sets $U_k^+$ and $U_k^-$ by (9) and (10).
        Compute the maximum decision value $f_k^{max}(\cdot)$ of samples in $U_k^+$ and the minimum decision value $f_k^{min}(\cdot)$ of samples in $U_k^-$.
        Compute the width ($W_k^+$) of each partition of samples in the subsets $U_k^+$ and the width ($W_k^-$) of each partition of samples in the subsets $U_k^-$ by (13).
        Compute the bounds ($H_k and L_k$) of each partition in $U_k^+$ and $U_k^-$ by (15).
        Use $H_k, L_k, W_k$ to partition $U_k^+$ and $U_k^-$ into $m = \frac{|U_k|}{a}$ partitions, respectively.
        Compute the set $P_k$ of non-empty partitions of two subsets $U_k^+$ and $U_k^-$.
        Choose the representative sample from each cluster by (11) and assign labels to them.
        Add the selected samples into the training set and update the unlabeled set.
        Retrain the classifiers with the updated training set using (2).
    end for
until stop criterion is satisfied
```

are retrained with the new labeled set. The process of the algorithm CLU-PBD is similar except for the step of selecting the informative samples outside the margin boundaries. As for the stop criterion, taking SVMs as the base model, we follow the model-specific stopping criterion[63] that the active learning process should stop when there are no unlabeled samples lying within the margin boundaries of classifiers, $w_k \cdot \phi(x) + b_k = 1$ and $w_k \cdot \phi(x) + b_k = -1$, since the samples (support vectors) within the boundaries support the classifier and the samples outside the boundaries will not alter the hyperplane when the obtained classifier is stable after iterations..

4. Experiment

4.1. Baselines and Metrics

In this section, we investigate the effectiveness of the proposed approaches (CLU-CBD and CLU-PBD) empirically. For comparison, two other active learning methods (VIO and Random) are used as baselines. The details of the two baselines are as follows. As for the metric, the average accuracy of classification is utilized.

- **VOI Method:**The value of information algorithm[64] utilizes a cross-task value of information criteria, in which the reward of a labeling assignment is propagated and measured over all relevant tasks.

$$VOI(Y, x) = \sum_{y} p(Y = y|x)R(p, Y = y, x), \tag{18}$$

  where R is the rewards function and we use $R(p, Y = y, x) = -\log_2 p(Y = y|x)$. This strategy is to select the sample which is the most uncertain over all tasks;

- **Random Method:**This method always randomly selects samples from the data, we then utilize random strategy to select samples from each task.

4.2. Data Sets and Settings

We evaluate the proposed multi-task active learning approaches by conducting experiments on 20 Newsgroups[1] data set which contains about $20,000$ documents taken from 20 newsgroups, Reuters-21578[2] which contains about 21578 documents from the Reuters newswire, and the Dermatology data[3] which is from the UCI data sets. Since these data sets are not originally designed for multi-task learning, referring to the operation in [65–67], we reorganize the data sets as follows.

20 Newsgroups. There are seven top categories in the 20 Newsgroups data set: ”alt”, ”comp”, ”misc”, ”rec”, ”sci”, ”soc” and ”talk”. These top categories are further divided into 20 sub-categories where each categories has 1000 samples. We define the tasks as top-category-classification problems. Referring to the operation in [68], we remove three categories ”alt”, ”soc” and ”misc”, since they are too small. The remaining four top categories(”comp”, ”rec”, ”sci”, ”talk”) are used to construct six multi-task learning sub-datasets by combining two of them as shown in Table 1. For example, to construct the

[1] http://qwone.com/ jason/20Newsgroups/

[2] http://kdd.ics.uci.edu/databases/reuters21578/

[3] http://archive.ics.uci.edu/ml/datasets/Dermatology

multi-task data set comp vs sci, we select one sub-category from ”comp” as positive sub-dataset and select one sub-category from ”sci” as negative sub-dataset for each task. In addition, each document is represented as a binary vector consisting of the 200 most discriminating words determined by Weka’s info-gain filter [69]. Using this split strategy ensures that tasks are relevant because they are under the same top categories. At the same time, the tasks are ensured to be different because they are drawn from different sub-categories.

Reuters-21578. There are five top categories among which ”orgs”, ”people” and ”places” are three big ones. Referring to the operation in [70], only three top categories are used in the experiments. Similar to the Setting in the 20 Newsgroups, three top categories are used to construct three data sets orgs vs people, orgs vs places and people vs places. Further, three data sets are reorganized into three multi-task sub-datasets with two tasks by dividing the sub-categories in each top categories. For example, the people vs places data set consists of two tasks. We randomly divide ”people” which has 267 sub-categories into two parts. Similarly, ”places” is also divided randomly into two parts. Documents in ”people” are considered as positive and documents in ”places” are considered as negative. Therefore, the two tasks are related since the positive classes and the negative classes belong to the same top categories, respectively. As done in the 20 Newsgroups, each document is represented as a binary vector consisting of the 200 most discriminating words.

Dermatology data. The Dermatology data which is used for the differential diagnosis of erythematous-squamous diseases, consists of 6 diseases with very little differences: psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. The 6 diseases are considered as six categories, which contain 112, 61, 72, 49, 52, 20 samples, respectively. This data set contains 34 attributes, 33 of which are linear valued and one of them is nominal. To divide the data set to a multi-task data set, five classification tasks get constructed. The purpose of each task is to diagnose one of six dermatological disease. For example, in the first classification task, the positive samples are taken from psoriasis and the negative samples are taken from lichen planus.

Each sub-dataset is randomly divided into two parts: the unlabeled training set(65%) and the labeled testing set(35%). 5% samples are randomly selected from the training set as an initial labeled set. The performance of each classifier is measured based on classification accuracy on the test set at the end of each iteration. For the parameter settings of the proposed methods, the value of clustering parameter h is set as $h = \frac{|U_k|}{a}$, where $|U_k|$ is the number of unlabeled samples in the training set for task k. The value of the partition parameter m is set as $m = \frac{|U_k|}{a}$. In the experiments, a is set as 40. The tradeoff parameter is fixed to $\mathcal{C} = 0.1$ in all experiments. To void the bias, we repeat the above experiment ten times and report the average accuracy.

Table 1: Description of the data sets

	Data set	Task number	Positive sub-dataset	Negative sub-dataset
1	comp vs. rec	4	comp.graphics	rec.autos
			comp.os.ms-windows.misc	rec.motorcycles
			comp.sys.ibm.pc.hardware	rec.sport.baseball
			comp.sys.mac.hardware	rec.sport.hockey
2	rec vs. sci	4	rec.autos	sci.crypt
			rec.motorcycles	sci.electronics
			rec.sport.baseball	sci.med
			rec.sport.hockey	sci.space
3	comp vs. sci	4	comp.os.ms-windows.misc	sci.med
			comp.sys.ibm.pc.hardware	sci.crypt
			comp.graphics	sci.space
			comp.sys.mac.hardware	sci.electronics
4	sci vs. talk	3	sci.crypt	talk.politics.guns
			sci.electronics	talk.politics.mideast
			sci.med	talk.politics.misc
5	talk vs. comp	3	talk.politics.mideast	comp.sys.ibm.pc.hardware
			talk.religion.misc	comp.graphics
			talk.politics.guns	comp.sys.mac.hardware
6	talk vs. rec	3	talk.religion.misc	rec.sport.hockey
			talk.politics.misc	rec.sport.baseball
			talk.politics.guns	rec.autos
7	orgs vs. people	2	part of samples in orgs	part of samples in people
			another part of samples in orgs	another part of samples in peo
8	people vs. places	2	part of samples in people	part of samples in places
			another part of samples in people	another part of samples in pla
9	orgs vs. places	2	part of samples in places	part of samples in orgs
			another part of samples in places	another part of samples in org
10	Dermatology	5	psoriasis	lichen planus
			lichen planus	seboreic dermatitis
			cronic dermatitis	pityriasis rosea
			seboreic dermatitis	pityriasis rubra pilaris
			pityriasis rosea	cronic dermatitis

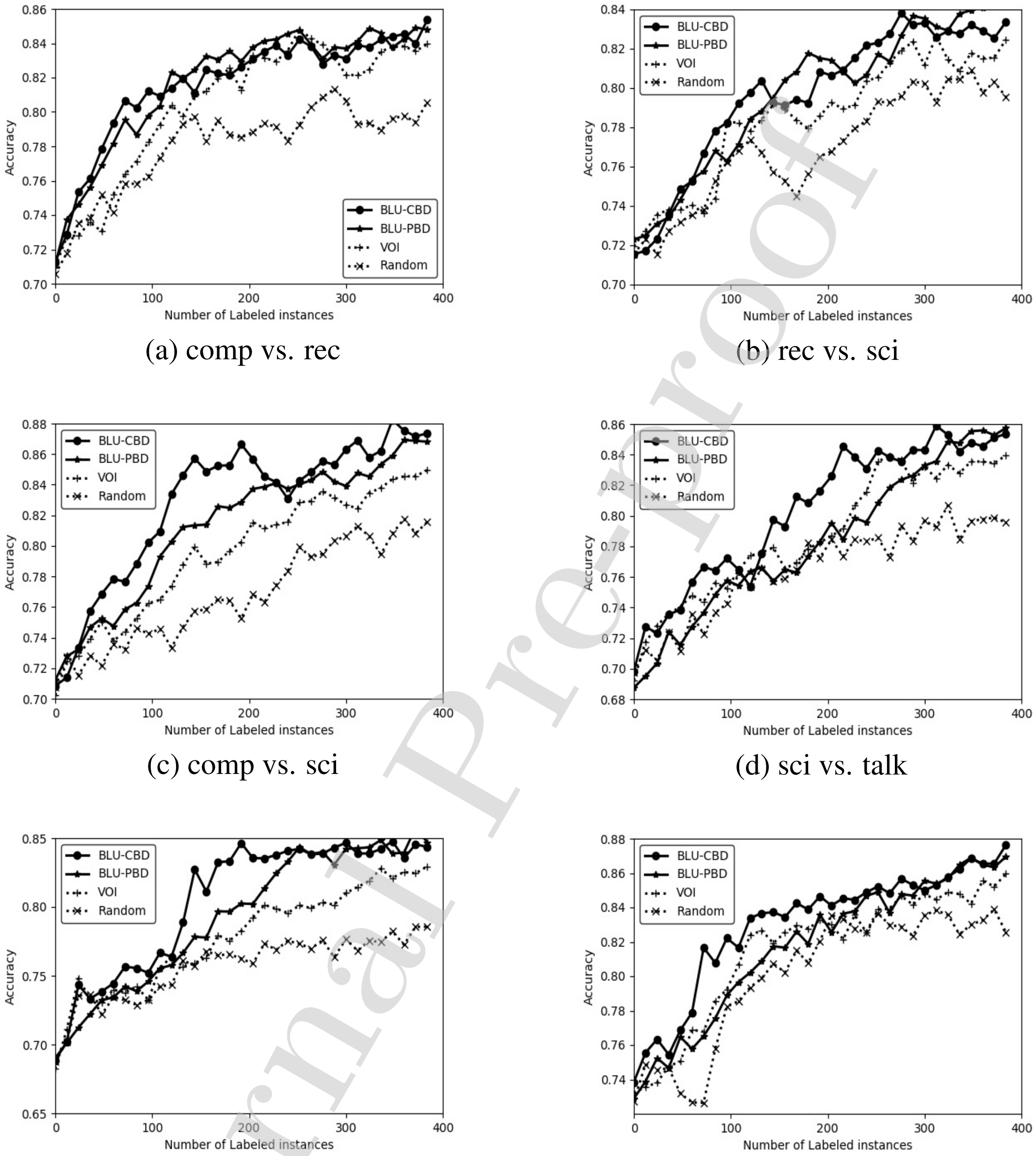
4.3. Performance Comparison

Table 2 presents the overall average classification accuracy on the ten data sets as the number of labeled samples is set as 50. It can be seen from the Table that the pro-

Table 2: Average classification accuracy on 10 sub-datasets when the number of labeled instances is 50.

Accuracy \ Method Sub-dataset	Random	VOI	CLU-PBD	CLU-CBD
comp vs. rec	0.7328	0.7564	0.7789	0.7886
rec vs. sci	0.7358	0.7386	0.7432	0.7486
comp vs. sci	0.7216	0.7537	0.7526	0.7623
sci vs. talk	0.7114	0.7334	0.7155	0.7379
talk vs. comp	0.7224	0.7313	0.7342	0.7415
talk vs. rec	0.7331	0.7562	0.7647	0.7756
orgs vs. people	0.7629	0.7684	0.7965	0.7739
orgs vs. places	0.7246	0.7548	0.7518	0.7784
people vs. places	0.7255	0.7383	0.7626	0.7649
Dermatology	0.6453	0.6513	0.6716	0.6794

posed CLU-CBD and CLU-PBD methods outperform random-based and VOI-based active learning methods in all, this is because the proposed methods extract the informative samples around the classifier and the representative examples, which can maintain the basic structure of the data, while Random and VOI methods do not explicitly extract these structure data. VOI method only extracts samples based on their total information value for all tasks. However, a sample with maximum total information value for all tasks dose not mean that it is valuable for each task. Random method always selects samples randomly from the data of each task, therefore, the selected samples are not always informative. Therefore, the two proposed methods perform better than VOI and Random methods. In addition, we find that the proposed CLU-CBD method performs better than CLU-PBD method for most data sets, since the micro-clustering method can always cluster the non-support vector data into a number of condense clusters, and the clusters can cover most of the data, however, the partition-based method splits the non-support vector data according to the distance to the hyperplane, and select one representative sample from each non-empty partition. Thus, the proposed CLU-CBD can select more representative samples than CLU-PBD method, and the former method outperforms the latter one in all. Furthermore, we observe that VOI method is better than Random method, since Random method always selects samples randomly for each task, and ignore some informative samples, this results the lowest performance compared with other methods.



(a) comp vs. rec

(b) rec vs. sci

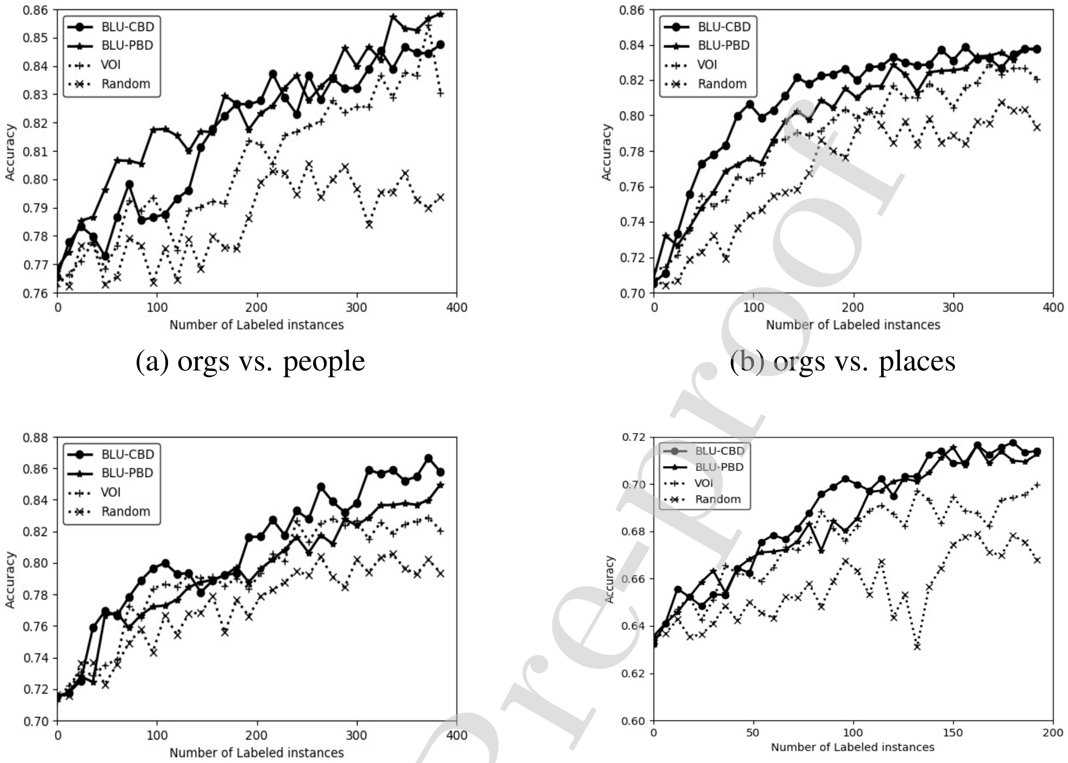
(c) comp vs. sci

(d) sci vs. talk

(d) talk vs. comp

(d) talk vs. rec

Figure 2: Comparative performance on 20 Newsgroups data.



(a) orgs vs. people

(b) orgs vs. places

(c) people vs. places

(d) dermatology

Figure 3: Comparative performance on Reuters-21578 data and Dermatology data.

4.4. Performance Variation as Labeled Samples Increase

In Figure 2-3, we present the variation of overall average classification accuracy on the ten data sets as the number of labeled samples increases. Figure 2 illustrates the results for the data set 1 to data set 6, and Figure 3 shows the results for data set 7 to data set 10. From the figures, we observe that as the number of labeled samples increases, the performance of all the methods increases in all, since more labeled samples can contain more data information and build more accurate multi-task classifiers. However, we can still find that the overall average accuracy of the proposed CLU-CBD and CLU-PBD methods can always yield higher performance than VOI and Random methods, since the proposed methods select local and global informative samples for each task at the same time, and this results in higher performance compared with other methods.

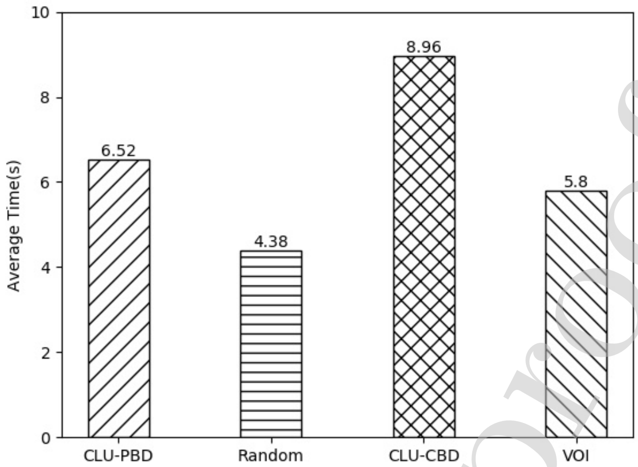


Figure 4: Average computational time

4.5. *Computational Time Comparison*

Figure 4 presents the average computational time required by the different methods. From this figure, one can see that the computational time required by Random method is less than other methods, because it selects samples randomly rather than strategically. The proposed CLU-CBD and CLU-PBD methods require more computational time than VOI method, since they adopt clustering and partition methods to avoid sample redundancy respectively. As the comparison of the proposed CLU-CBD and CLU-PBD methods, the CLU-CBD method costs more time than the CLU-PBD method, since the CLU-CBD method utilizes micro-cluster method to split the data, and requires more time than the partition method.

5. Conclusion

In this paper, we propose two different active learning methods and incorporate them with multi-task SVM to cope with multi-task classification problems. We generalize the proposed methods (CLU-CBD and CLU-PBD) based on CLU in the uncertainty step, and CBD and PBD in the diversity step to multi-task problems. CLU-CBD strategy exploits the support vectors of each hyperplane for queries in the uncertainty step. Furthermore, by means of micro-kernel k-means clustering, it mines the informative samples based on the distribution of the unlabeled samples in the diversity step. CLU-PBD method also exploits the support vectors of each hyperplane for queries in the uncertainty step, and mines the informative samples based on the distribution of the unlabeled samples by means of partition in the diversity step. We carry out the process of active learning simultaneously in each task to ensure that the classifier for each task is optimized. In the experimental analysis, the proposed techniques get compared with other active learning methods adopted in multi-task classification on three data sets. The experimental results show that the proposed methods can take advantage of the informative samples to improve the classification accuracy compared with other active learning methods.

In future work, we plan to both extend the experimental comparison to other active learning methods and extend the proposed methods to the multi-class setting. In addition, we plan to apply the proposed active learning methods in the data stream environments.

References

[1] Tony Jebara. Multi-task feature and kernel selection for svms. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, 2004.

[2] You Ji and Shiliang Sun. Multitask multiclass support vector machines: Model and experiments. *Pattern Recognition*, 46(3):914–924, 2013.

[3] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 109–117, 2004.

[4] Deyu Zhou, Lei Miao, and Yulan He. Position-aware deep multi-task learning for drug-drug interaction extraction. *Artificial Intelligence in Medicine*, 87:1–8, 2018.

[5] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S. Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(5):1167–1181, 2018.

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. *CoRR*, abs/1607.05369, 2016.

[7] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Trans. Image Processing*, 23(8):3656–3670, 2014.

[8] Xu Cheng, Nijun Li, Tongchi Zhou, Zhenyang Wu, and Lin Zhou. Multi-task object tracking with feature selection. *IEICE Transactions*, 98-A(6):1351–1354, 2015.

[9] Zhangjian Ji and Weiqiang Wang. Robust object tracking via multi-task dynamic sparse model. In *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, pages 393–397, 2014.

[10] Xiyan He, Gilles Mourot, Didier Maquin, José Ragot, Pierre Beauseroy, André Smolarz, and Edith Grall-Maës. Multi-task learning with one-class SVM. *Neurocomputing*, 133:416–426, 2014.

[11] Liyun Lu, Qiang Lin, Huimin Pei, and Ping Zhong. The als-svm based multi-task learning classifiers. *Appl. Intell.*, 48(8):2393–2407, 2018.

[12] Han-Tai Shiao and Vladimir Cherkassky. Implementation and comparison of svm-based multi-task learning methods. In *The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012*, pages 1–7, 2012.

[13] Jean Baptiste Faddoul, Boris Chidlovskii, Rémi Gilleron, and Fabien Torre. Learning multiple tasks with boosted decision trees. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part I*, pages 681–696, 2012.

[14] Qing Wang, Liang Zhang, Mingmin Chi, and Jiankui Guo. Mtforest: Ensemble decision trees based on multi-task learning. In *ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, Proceedings*, pages 122–126, 2008.

[15] Bin Huang, Dengfeng Ke, Hao Zheng, Bo Xu, Yanyan Xu, and Kaile Su. Multi-task learning deep neural networks for speech feature denoising. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2464–2468, 2015.

[16] Simeng Yue and Seiichi Ozawa. A sequential multi-task learning neural network with metric-based knowledge transfer. In *11th International Conference on Machine Learning and Applications, ICMLA, Boca Raton, FL, USA, December 12-15, 2012. Volume 1*, pages 671–674, 2012.

[17] Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. Multi-task deep neural network for multi-label learning. In *IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013*, pages 2897–2900, 2013.

[18] Xiao Chu, Wanli Ouyang, Wei Yang, and Xiaogang Wang. Multi-task recurrent neural network for immediacy prediction. In *2015 IEEE International Conference on*

Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 3352–3360, 2015.

[19] Chengtao Li, Jun Zhu, and Jianfei Chen. Bayesian max-margin multi-task learning with data augmentation. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 415–423, 2014.

[20] Keelin Greenlaw, Elena Szefer, Jinko Graham, Mary Lesperance, and Farouk S. Nathoo. A bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics*, 33(16):2513–2522, 2017.

[21] Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2004–2012, 2013.

[22] Jian Wu, Anqian Guo, Victor S. Sheng, Pengpeng Zhao, and Zhiming Cui. An active learning approach for multi-label image classification with sample noise. *IJPRAI*, 32(3):1–23, 2018.

[23] Min Wang, Fan Min, Zhi-Heng Zhang, and Yan-Xue Wu. Active learning through density clustering. *Expert Syst. Appl.*, 85:305–317, 2017.

[24] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 917–926, 2009.

[25] Begüm Demir, Luca Minello, and Lorenzo Bruzzone. Definition of effective training sets for supervised classification of remote sensing images by a novel cost-sensitive active learning method. *IEEE Trans. Geoscience and Remote Sensing*, 52(2):1272–1284, 2014.

[26] Husheng Guo and Wenjian Wang. An active learning-based SVM multi-class classification model. *Pattern Recognition*, 48(5):1577–1597, 2015.

[27] Saad Mohamad, Moamar Sayed Mouchaweh, and Abdelhamid Bouchachia. Active learning for classifying data streams with unknown number of classes. *Neural Networks*, 98:1–15, 2018.

[28] Erelcan Yanik and Tevfik Metin Sezgin. Active learning for sketch recognition. *Computers & Graphics*, 52:93–105, 2015.

[29] Suntae Kim, Jintae Kim, and Sooyong Park. An active learning framework for object-oriented analysis and design. *Comp. Applic. in Engineering Education*, 20(3):400–409, 2012.

[30] John M. Carroll and Robert L. Mack. Metaphor, computing systems, and active learning. *International Journal of Man-Machine Studies*, 22(1):39–57, 1985.

[31] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 746–751, 2005.

[32] David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[33] Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, and Thomas Hopkins. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6:589–613, 2005.

[34] Maria-Florina Balcan, Andrei Z. Broder, and Tong Zhang. Margin based active learning. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13-15, 2007, Proceedings*, pages 35–50, 2007.

[35] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[36] Saeed Majidi and Gregory R. Crane. Committee-based active learning for dependency parsing. In *Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings*, pages 442–445, 2013.

[37] Handing Wang, Yaochu Jin, and John Doherty. Committee-based active learning for surrogate-assisted particle swarm optimization of expensive problems. *IEEE Trans. Cybernetics*, 47(9):2664–2677, 2017.

[38] Komei Sugiura, Naoto Iwahashi, Hideki Kashioka, and Satoshi Nakamura. Active learning of confidence measure function in robot language acquisition framework. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*, pages 1774–1779, 2010.

[39] Oscar Gabriel Reyes Pupo, Carlos Morell, and Sebastián Ventura. Effective active learning strategy for multi-label learning. *Neurocomputing*, 273:494–508, 2018.

[40] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992.*, pages 287–294, 1992.

[41] Michael Davy and Saturnino Luz. An adaptive pre-filtering technique for error-reduction sampling in active learning. In *Workshops Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 682–691, 2008.

[42] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *J. Artif. Intell. Res.*, 4:129–145, 1996.

[43] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 441–448, 2001.

[44] Jinbo Bi, Tao Xiong, Shipeng Yu, Murat Dundar, and R. Bharat Rao. An improved multi-task learning approach with applications in medical diagnosis. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I*, pages 117–132, 2008.

[45] Seiichi Ozawa, Asim Roy, and Dmitri Roussinov. A multitask learning model for online pattern recognition. *IEEE Trans. Neural Networks*, 20(3):430–445, 2009.

[46] Sahil Sharma and Balaraman Ravindran. Online multi-task learning using active sampling. *CoRR*, abs/1702.06053, 2017.

[47] Zhifeng Hao, Yibang Ruan, Yanshan Xiao, and Bo Liu. A multi-task-based classification framework for multi-instance distance metric learning. *Neurocomputing*, 275:418–429, 2018.

[48] Jun Zhu, Ning Chen, and Eric P. Xing. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 1620–1628, 2011.

[49] Tony Jebara. Multi-task feature and kernel selection for svms. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, 2004.

[50] Han-Tai Shiao and Vladimir Cherkassky. Implementation and comparison of svm-based multi-task learning methods. In *The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012*, pages 1–7, 2012.

[51] Yaran Chen, Dongbin Zhao, Le Lv, and Qichao Zhang. Multi-task learning for dangerous object detection in autonomous driving. *Inf. Sci.*, 432:559–571, 2018.

[52] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2873–2879, 2016.

[53] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pages 94–108, 2014.

[54] Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 599–606, 2010.

[55] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 1015–1022, 2007.

[56] Michael Pearce and Jürgen Branke. Continuous multi-task bayesian optimisation with correlation. *European Journal of Operational Research*, 270(3):1074–1085, 2018.

[57] Haiqin Yang, Irwin King, and Michael R. Lyu. Multi-task learning for one-class classification. In *International Joint Conference on Neural Networks, IJCNN 2010, Barcelona, Spain, 18-23 July, 2010*, pages 1–8, 2010.

[58] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 475–482, 2005.

[59] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*, pages 41–48, 1993.

[60] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

[61] Martijn Van Otterlo, Joint Work Kristian Kersting, Luc De, Raedt Freiburgleuven, Workshop March, and Martijn Otterlo. *Markov Decision Processes*. John Wiley and Sons, 1993.

[62] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Publications of the American Statistical Association*, 101(476):1566–1581, 2006.

[63] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 839–846, 2000.

[64] Yi Zhang. Multi-task active learning with output constraints. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, 2010.

[65] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 193–200, 2007.

[66] Samir Al-Stouhi and Chandan K. Reddy. Multi-task clustering using constrained symmetric non-negative matrix factorization. In *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pages 785–793, 2014.

[67] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 253–261, 2013.

[68] Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, pages 342–357, 2008.

[69] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[70] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 210–219, 2007.

Yanshan Xiao received the Ph.D. degree in computer science from the Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia, in 2011. She is with the Faculty of Computer, Guangdong University of Technology. Her research interests include data mining and machine learning. She has published papers on IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, Knowledge and Information Systems, and International Joint Conferences on Artificial Intelligence (IJCAI).

Zheng Chang is pursuing a master's degree at the school of Computers, Guangdong University of Technology, China. His research interests include machine learning and data mining.

Bo Liu is with the Faculty of Automation, Guangdong University of Technology. His research interests include machine learning and data mining. He has published papers on IEEE Transactions on Neural Networks, IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, IEEE International Conference on Data Mining (ICDM), SIAM International Conference on Data Mining (SDM) and ACM International Conference on Information and Knowledge Management (CIKM).