Diversifying Convex Transductive Experimental Design for Active Learning

Lei Shi[1,2] and **Yi-Dong Shen**[1]*
[1]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, Beijing 100190, China
{shilei,ydshen}@ios.ac.cn

Abstract

Convex Transductive Experimental Design (CTED) is one of the most representative active learning methods. It utilizes a data reconstruction framework to select informative samples for manual annotation. However, we observe that CTED cannot well handle the diversity of selected samples and hence the set of selected samples may contain mutually similar samples which convey similar or overlapped information. This is definitely undesired. Given limited budget for data labeling, it is desired to select informative samples with complementary information, i.e., similar samples are excluded. To this end, we proposes Diversified CTED by seamlessly incorporating a novel and effective diversity regularizer into CTED, ensuring the selected samples are diverse. The involvement of the diversity regularizer leads the optimization problem hard to solve. We derive an effective algorithm to solve an equivalent problem which is easier to optimize. Extensive experimental results on several benchmark data sets demonstrate that Diversified CTED significantly improves CTED and consistently outperforms the state-of-the-art methods, verifying the effectiveness and advantages of incorporating the proposed diversity regularizer into CTED.

1 Introduction

In many machine learning tasks, we need to collect the training data and manually annotate them by domain experts. This process is usually time consuming and expensive. Active learning [Settles, 2009] is a machine learning technique that selects the most informative samples for labeling and uses them as training data. It has been widely explored in the machine learning community for its capability of reducing human annotation effort.

Convex Transductive Experimental Design (CTED) [Yu *et al.*, 2008] is one of the most representative active learning methods and has received increasing attention in recent years. It uses a data reconstruction framework to select informative samples for labeling, where the informativeness of each sample is measured by its capacity to reconstruct the target data set. Based on the data reconstruction framework, several methods were developed. [Zhen and Yeung, 2010] proposed supervised experimental design, which can leverage label information if it is available. [Nie *et al.*, 2013; Zhu and Fan, 2015] employed robust loss functions to measure the reconstruction error to develop methods which are insensitive to outliers. [Cai and He, 2012] incorporated manifold structure into this framework via manifold adaptive kernels. [Hu *et al.*, 2013] performed active learning via neighbourhood reconstruction to select samples by exploring the local data structure. These methods have achieved promising performance in text classification [Yu *et al.*, 2008; Cai and He, 2012] and other multimedia data classification tasks [Nie *et al.*, 2013; Hu *et al.*, 2013; Zhu and Fan, 2015].

Observe that the data reconstruction framework assigns each sample a score, which indicates the sample's capacity to reconstruct the target data set, and all the samples are ranked based on these scores. Then the top ranked ones are selected for labeling. Similar samples may get similar ranking scores, because these similar ones have similar capacity for data reconstruction. As a result, the set of selected (i.e., top ranked) samples may well contain samples which are mutually similar. This is definitely undesired. Since the process of data labeling is usually time consuming and expensive, the budget for data labeling is always limited. To maximize our benefit, it is desired to select those informative samples containing complementary information, i.e., highly similar ones are excluded.

The above analysis motivates us to study the diversity problem with existing data reconstruction based active learning methods. In particular, we propose to enhance CTED with diversity mechanism by imposing a diversity regularizer over sample selection. The diversity regularizer makes use of a similarity matrix among samples to ensure that if two samples are informative but highly similar, only one of them gets a high ranking score. A main challenge in applying the diversity regularizer is to define a good similarity matrix. One direct way to obtain the similarity matrix is to pre-define it using the original data. Then it is fixed and used as an input to the diversity regularizer. If the pre-defined similarity matrix is not good, the effectiveness of the diversify regularizer will be

*Corresponding author

limited. To obtain a reliable similarity matrix and seamlessly integrate diversity into CTED, we embed similarity matrix learning into CTED by leveraging a data representation extracted from the reconstruction matrix of CTED. As we will discuss later, this data representation reflects the role of each sample in the data reconstruction process. Therefore, similarity matrix built on this representation can well characterize the similarities among samples in terms of their capacity for reconstruction. The involvement of the diversity regularizer makes the obtained optimization problem hard to solve. We find an equivalent problem which is easier to optimize and derive an alternating minimization procedure to solve it. We perform extensive experiments on several benchmark data sets. Experimental results based on two classifiers all demonstrate that, with diversification, our method (i.e., Diversified CTED) significantly improves CTED and consistently outperforms the state-of-the-art methods in the literature.

2 Related Work

The goal of active learning is to label as little data as possible, to achieve a certain classification performance, therefore saving considerable annotation cost for training a good learner [Settles, 2009]. One representative active learning algorithm is to select samples with maximum uncertainty of labels measured by the distance from the classification boundary [Tong and Koller, 2002; Lewis and Catlett, 1994; Balcan *et al.*, 2007; Yang *et al.*, 2014]. Another popular approach is query by committee, where a number of distinct classifiers are generated and a sample having the most disagreement among these classifiers in predicting the label is selected for labeling [Freund *et al.*, 1997; Seung *et al.*, 1992]. When selecting samples, these methods need a pretrained classifier, which means they need some initially labeled data. Another line of research works aim to select samples in unsupervised setting. Clustering based active learning methods were proposed in [Nguyen and Smeulders, 2004; Nie *et al.*, 2012]. [Chattopadhyay *et al.*, 2012; 2013] proposed an active learning method based on Maximum Mean Discrepancy (MMD)[Borgwardt *et al.*, 2006], with the goal of minimizing the difference in the marginal probability distribution between the selected samples and remaining ones. Data reconstruction based active learning methods, such as CTED [Yu *et al.*, 2008] and ARSS [Zhu and Fan, 2015], are also typical methods in this category.

In this paper, our focus is on solving the diversity issue of the data reconstruction based active learning methods. Particularly, we solve this problem on the base of CTED and show that, with diversification, the sample selection power of CTED can be significantly improved. The developed techniques can also be used to other data reconstruction based active learning methods.

3 Preliminaries

We summarize the notations and the definition of norms used in this paper. Matrices are written as boldface uppercase letters and vectors are written as boldface lowercase letters. For an arbitrary matrix $\mathbf{M}$, we denote its i-th row, j-th column and (i,j)-th entry as $\mathbf{m}^i$, $\mathbf{m}_j$ and m_{ij}, respectively. The ℓ_p-norm of the vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $||\mathbf{v}||_p = (\sum_{i=1}^n |v_i|^p)^{\frac{1}{p}}$. The Frobenius norm of the matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ is defined as $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m m_{ij}^2}$. $\mathbf{M}^T$ is the transpose of $\mathbf{M}$ and $\mathrm{Tr}(\mathbf{M})$ is the trace of $\mathbf{M}$. We use $\mathbf{I}$ to denote the identity matrix with proper size.

Denote $\mathbf{X} \in \mathbb{R}^{d\times n}$ as an unlabeled data set, where $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ is the i-th sample. The goal of active learning is to select a subset $\mathbf{Z} \in \mathbb{R}^{d\times m}$ $(m < n)$ from $\mathbf{X}$, such that the selected samples can improve the classifier the most if they are labeled and added to the training set. Since our work is based on Convex Transductive Experimental Design (CTED), we first review CTED in the next subsection.

3.1 Convex Transductive Experimental Design

Transductive Experimental Design (TED) [Yu *et al.*, 2006] aims at selecting a subset $\mathbf{Z} \in \mathbb{R}^{d\times m}$ from the original unlabeled data set $\mathbf{X} \in \mathbb{R}^{d\times n}$ such that a function f trained on $\mathbf{Z}$ has the smallest predictive variance on a given testing set. The optimization problem of TED is

$$\begin{aligned} \max_{\mathbf{Z}} \quad & \mathrm{Tr}(\mathbf{X}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z}+\mu\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{X}) \\ s.t. \quad & \mathbf{Z} \subset \mathbf{X}, |\mathbf{Z}| = m. \end{aligned} \tag{1}$$

where $|\mathbf{Z}| = m$ means that $\mathbf{Z}$ contains m samples and μ is a tuning parameter. In order to solve the NP-hard optimization problem in Eq. (1), [Yu *et al.*, 2006] proposed a sequential algorithm, which selects one sample each time. The obtained result is suboptimal [Yu *et al.*, 2008]. To tackle this issue, Convex Transductive Experimental Design (CTED) was further proposed in [Yu *et al.*, 2008]. The optimization problem of CTED is

$$\begin{aligned} \min_{\mathbf{A},\mathbf{b}} \quad & ||\mathbf{X}-\mathbf{X}\mathbf{A}||_F^2 + \sum_{i=1}^n \frac{\sum_{j=1}^n a_{ij}^2}{b_i} + \gamma||\mathbf{b}||_1 \\ s.t. \quad & b_i \geq 0, i = 1,2,...,n \end{aligned} \tag{2}$$

where a_{ij} is the (i,j)-th entry of $\mathbf{A}$, and γ is a nonnegative trade-off parameter. As shown in Eq. (2), CTED utilizes a data reconstruction framework to select informative samples for labeling. The matrix $\mathbf{A}$ contains reconstruction coefficients and $\mathbf{b}$ is the sample selection vector. The ℓ_1-norm makes $\mathbf{b}$ to be sparse. In this way, the values corresponding to less informative samples tend to be zero. Large value of b_i indicates the i-th sample has a large impact on the reconstruction. After solving the optimization problem in (2), CTED ranks all samples based on $b_i (i = 1,2,...,n)$ in descending order, and select the top ranked ones for labeling.

Under this data reconstruction framework, several methods have been proposed [Cai and He, 2012; Nie *et al.*, 2013; Zhen and Yeung, 2010; Zhu and Fan, 2015; Hu *et al.*, 2013]. For example, [Cai and He, 2012] incorporated manifold information into this framework. [Zhu and Fan, 2015] employed robust loss functions to measure reconstruction error to handle outliers in data. These methods have shown effectiveness in text classification and other multimedia data classification tasks. All these methods select informative samples via a same mechanism, i.e., assigning ranking scores to samples and selecting top ranked ones.

As discussed above, similar samples may get similar ranking scores, since these similar ones have similar capacity for reconstruction. As a result, the set of top ranked samples may well contain similar samples which convey overlapped information. Given limited data labeling budget, if we can exclude mutually similar samples and select ones with complementary information, the performance of these methods could be improved. In this paper, we develop techniques to solve this problem based on CTED.

4 Diversifying Convex Transductive Experimental Design

In this section, we propose to diversify the selected samples to improve CTED, i.e., highly similar samples are excluded. We first introduce a similarity matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ to model the pairwise similarities among all the samples, such that larger value of s_{ij} means higher similarity between the i-th sample and the j-th one, and smaller value indicates their similarity is lower. Intuitively, if two samples are highly similar to each other, they tend to convey very similar information. In this case, we intend the two samples not to be selected together. To realize this, we enhance CTED with a diversity regularizer by extending Eq. (2) to the following optimization problem

$$\begin{aligned} \min_{\mathbf{A},\mathbf{b}} \quad & ||\mathbf{X} - \mathbf{X}\mathbf{A}||_F^2 + \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} a_{ij}^2}{b_i} + \gamma ||\mathbf{b}||_1 + \alpha \mathbf{b}^T \mathbf{S} \mathbf{b} \\ s.t. \quad & b_i \geq 0, i = 1, 2, ..., n \end{aligned} \tag{3}$$

From Eq. (3), we can see that if s_{ij} is large (i.e., the i-th sample and the j-th one are highly similar), b_i and b_j cannot be large at the same time. This constraint guarantees that highly similar samples would not have higher scores in sample selection at the same time.

The effectiveness of the diversity regularizer in Eq. (3) highly relies on the quality of the similarity matrix $\mathbf{S}$. Properly defining a good $\mathbf{S}$ is the key for improving CTED by encouraging diversity of selected samples. A direct way to define $\mathbf{S}$ is to pre-calculate it based on the original data. Note that once the similarity matrix is defined, it is fixed in the following sample selection step. If the fixed similarity matrix is not well defined, the effectiveness of the diversity regularizer will be limited. To obtain a faithful similarity matrix and seamlessly incorporate diversity into CTED, we propose to embed similarity matrix learning into CTED via properly leveraging the special structure of the data reconstruction framework.

We observe that, the data reconstruction matrix $\mathbf{A}$ encodes relations among samples. Let us take a look at the i-th row of $\mathbf{A}$, i.e., $\mathbf{a}^i$, which corresponds to the i-th sample. a_{ij} means the j-th sample's reconstruction coefficient based on the i-th sample. And $\mathbf{a}^i$ encodes the reconstruction coefficients of all the samples based on the i-th one. The i-th row of $\mathbf{A}$ (i.e., $\mathbf{a}^i$) can be treated as a data representation which reflects the role of the i-th sample for data reconstruction. If i-th and j-th samples are similar in terms of their capacity for reconstruction, they tend to have similar $\mathbf{a}^i$ and $\mathbf{a}^j$. Therefore, we use $s_{ij} = \frac{\mathbf{a}^i (\mathbf{a}^j)^T}{||\mathbf{a}^i||_2 ||\mathbf{a}^j||_2}$ to characterize the similarity between the i-th sample and j-th one. This similarity measure directly reflects the similarity relationship among samples in terms of their ability for reconstruction. Based on the above analysis, we define a novel and effective diversity regularizer and use it to regularize CTED. We get the following optimization problem of Diversified CTED (DCTED for short)

$$\begin{aligned} \min_{\mathbf{A},\mathbf{b}} \quad & h(\mathbf{A}, \mathbf{b}) = ||\mathbf{X} - \mathbf{X}\mathbf{A}||_F^2 + \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} a_{ij}^2}{b_i} + \gamma ||\mathbf{b}||_1 \\ & + \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\mathbf{a}^i (\mathbf{a}^j)^T}{||\mathbf{a}^i||_2 ||\mathbf{a}^j||_2} b_i b_j \\ s.t. \quad & b_i \geq 0, i = 1, 2, ..., n \end{aligned} \tag{4}$$

Note that the similarity matrix in Eq. (4) is adaptively updated in the learning process. We do not need to predefine a similarity matrix. This nice property makes our model very practical in real world applications. Our empirical studies also suggest that this diversity term is more effective than the one with a predefined similarity matrix. The involvement of the diversity regularizer makes the optimization problem in Eq. (4) hard to solve. In the next section, we derive an effective algorithm to solve it.

5 Optimization

The optimization problem in Eq. (4) involves two groups of variables, i.e., $\mathbf{A}$ and $\mathbf{b}$. It is hard to solve due to the complex structure of the diversity term. We first reformulate this problem. Let $\mathbf{A} = \mathrm{diag}(\mathbf{s})\widehat{\mathbf{A}}$, where $\mathrm{diag}(\mathbf{s})$ is a diagonal matrix whose diagonal elements are formed by the vector $\mathbf{s}$. The i-th element (i.e., s_i) of $\mathbf{s}$ is the ℓ_2-norm of the i-th row of $\mathbf{A}$. Therefore, $||\widehat{\mathbf{a}}^i||_2 = 1$. The problem in Eq. (4) can be reformulated as

$$\begin{aligned} \min_{\mathbf{A},\mathbf{b},\mathbf{s}} \quad & k(\widehat{\mathbf{A}}, \mathbf{s}, \mathbf{b}) = ||\mathbf{X} - \mathbf{X}\mathrm{diag}(\mathbf{s})\widehat{\mathbf{A}}||_F^2 + \sum_{i=1}^{n} \frac{s_i^2 \sum_{j=1}^{n} \widehat{a}_{ij}^2}{b_i} \\ & + \gamma ||\mathbf{b}||_1 + \alpha \mathbf{b}^T \widehat{\mathbf{A}} \widehat{\mathbf{A}}^T \mathbf{b} \\ s.t. \quad & b_i \geq 0, s_i \geq 0, ||\widehat{\mathbf{a}}^i||_2 = 1, i = 1, 2, ..., n \end{aligned} \tag{5}$$

Note that we can replace $\sum_{j=1}^{n} \widehat{a}_{ij}^2$ as 1. We keep it in Eq .(5) to better illustrate the relationship between the optimization problems in Eq. (4) and (5).

Theorem 1 *The optimization problem in Eq.* (5) *is equivalent to the one in Eq.* (4).

Proof 1 *If* $(\boldsymbol{A}^*, \boldsymbol{b}^*)$ *is the solution of the problem in Eq.* (4)*, then* $h(\boldsymbol{A}^*, \boldsymbol{b}^*) \leq h(\boldsymbol{A}, \boldsymbol{b})$*. Let* $\boldsymbol{A}^* = \mathrm{diag}(\boldsymbol{s}^*)\widehat{\boldsymbol{A}}^*$*, where* s_i^* *is the* ℓ_2*-norm of the* i*-th row of* $\boldsymbol{A}^*$*, we get* $k(\widehat{\boldsymbol{A}}^*, \boldsymbol{s}^*, \boldsymbol{b}^*) = h(\boldsymbol{A}^*, \boldsymbol{b}^*)$*. For* $(\widehat{\boldsymbol{A}}, \boldsymbol{s}, \boldsymbol{b})$*, it is easy to get* $k(\widehat{\boldsymbol{A}}, \boldsymbol{s}, \boldsymbol{b}) = h(\boldsymbol{A}, \boldsymbol{b}) \geq h(\boldsymbol{A}^*, \boldsymbol{b}^*) = k(\widehat{\boldsymbol{A}}^*, \boldsymbol{s}^*, \boldsymbol{b}^*)$*. This means* $(\widehat{\boldsymbol{A}}^*, \boldsymbol{s}^*, \boldsymbol{b}^*)$ *is the solution of the problem in Eq.* (5)*. Therefore, if we have the solution of the problem in Eq.* (4)*, we can readily to get the solution of the problem in Eq.* (5)*. Similarly, if we have the solution of Eq.* (5)*, we can also readily to get the solution of Eq.* (4)*. Therefore, the optimization problem in Eq.* (5) *is equivalent to the one in Eq.* (4)*.*

In next, we derive an alternating minimization algorithm to solve the problem in Eq. (5).

5.1 Update $\widehat{\mathbf{A}}$

The subproblem w.r.t. $\widehat{\mathbf{A}}$ is

$$\begin{aligned} \min_{\widehat{\mathbf{A}}} \quad & f(\widehat{\mathbf{A}}) = ||\mathbf{X} - \mathbf{X}\text{diag}(\mathbf{s})\widehat{\mathbf{A}}||_F^2 + \alpha \text{Tr}(\widehat{\mathbf{A}}^T \mathbf{b}\mathbf{b}^T \widehat{\mathbf{A}}) \\ s.t. \quad & ||\widehat{\mathbf{a}}^i||_2 = 1, i = 1, 2, ..., n \end{aligned} \tag{6}$$

We derive an algorithm to update $\widehat{\mathbf{A}}$ based on projected gradient descent to decrease the objective value of Eq. (6). Denote the gradient of $f(\widehat{\mathbf{A}})$ as $\nabla f(\widehat{\mathbf{A}})$. Then, $\widehat{\mathbf{A}}$ is updated as

$$\widehat{\mathbf{A}} = \widehat{\mathbf{A}} - \mu \nabla f(\widehat{\mathbf{A}}) \tag{7}$$

where μ is the step size. Since $\widehat{\mathbf{A}}$ should satisfies $||\widehat{\mathbf{a}}^i||_2 = 1$, we achieve this by projecting each row of $\widehat{\mathbf{A}}$ to the unit sphere. We need to search the step size to ensure the objective value is non-increasing under the updating rule of $\widehat{\mathbf{A}}$.

5.2 Update b

The subproblem w.r.t. $\mathbf{b}$ is

$$\begin{aligned} \min_{\mathbf{b}} \quad & \sum_{i=1}^{n} \frac{s_i^2}{b_i} + \gamma ||\mathbf{b}||_1 + \alpha \mathbf{b}^T \widehat{\mathbf{A}} \widehat{\mathbf{A}}^T \mathbf{b} \\ s.t. \quad & b_i \geq 0, i = 1, 2, ..., n \end{aligned} \tag{8}$$

The optimization problem is convex w.r.t. b_i when the other elements of $\mathbf{b}$ fixed. The subproblem w.r.t. b_i is

$$\begin{aligned} \min_{b_i} \quad & \frac{s_i^2}{b_i} + b_i(\gamma + 2\alpha \sum_{j \neq i}^{n} m_{ij} b_j) + \alpha m_{ii} b_i^2 \\ s.t. \quad & b_i \geq 0 \end{aligned} \tag{9}$$

where m_{ij} is the (i, j)-th entry of $\mathbf{M} = \widehat{\mathbf{A}} \widehat{\mathbf{A}}^T$. Set the derivative w.r.t. b_i as zero, we get

$$2\alpha m_{ii} b_i^3 + (\gamma + 2\alpha \sum_{j \neq i}^{n} m_{ij} b_j) b_i^2 - s_i^2 = 0 \tag{10}$$

It is easy to check that there exists positive solution of Eq. (10), which means the constraint $b_i \geq 0$ can be satisfied. We choose the positive solution to update b_i.

5.3 Update s

The subproblem w.r.t. $\mathbf{s}$ is

$$\begin{aligned} \min_{\mathbf{s}} \quad & ||\mathbf{X} - \mathbf{X}\text{diag}(\mathbf{s})\widehat{\mathbf{A}}||_F^2 + \sum_{i=1}^{n} \frac{s_i^2}{b_i} \\ s.t. \quad & s_i \geq 0, i = 1, 2, ..., n \end{aligned} \tag{11}$$

The optimization problem is convex w.r.t. s_i when other elements of $\mathbf{s}$ fixed. We can find s_i^* which leads the derivative w.r.t. s_i to be zero. However, s_i^* may be negative. Therefore, we update s_i as

$$s_i = max\{s_i^*, 0\} \tag{12}$$

Algorithm 1 The Optimization Algorithm for Diversified CTED

Input: $\mathbf{X}$, γ, and α
Output: Sort all the samples according to b_i $(i = 1, 2, ..., n)$ and select the top m ranked ones

```
Initialize Â, s and b randomly
repeat
    Update Â by Projected Gradient Descent
    Update b by solving Eq. (10)
    Update s by Eq. (12)
until Convergence
```

We summarize the whole procedure to solve the optimization problem in Eq. (5) as Algorithm 1. Since updating $\widehat{\mathbf{A}}$ decreases the objective value of the optimization problem in Eq. (5) and solving $\mathbf{b}$ and $\mathbf{s}$ also decrease it, therefore the updating rules decrease the objective value of the optimization problem in Eq. (5). As the optimization problem in Eq. (5) is non-negative and lower bounded, therefore Algorithm 1 converges. Since the optimization problem in Eq. (5) is not convex w.r.t. all the variables simultaneously, different initializations of these variables may lead Algorithm 1 converges to different values. For simplicity, we initialize these variables randomly. Our empirical studies show that Algorithm 1 works well with this simple initialization strategy.

6 Experiments

Following a same experimental protocol in [Yu *et al.*, 2008], we perform classification experiments on five benchmark data sets to demonstrate the effectiveness of the proposed method (i.e., Diversified CTED) and give analysis on the experimental results. We also perform experiments to study the effects of the parameters.

6.1 Datasets

We conduct the experiments on 5 publicly available data sets, including 2 digit recognition data sets (i.e., USPS [Wu and Schölkopf, 2006] and MNIST [Liu *et al.*, 2010]), 2 text data sets (i.e., WebKB [Wang *et al.*, 2011] and Newsgroup [Yu *et al.*, 2005]) and one face data set (i.e., ORL [Cai *et al.*, 2006]). The details of these data sets are summarized in Table 1.

Table 1: Summary of Data Sets

Dataset	Size	Dimensions	Classes
USPS	3082	256	4
MNIST	10000	784	10
NewsGroup	3970	8014	4
WebKB	4199	1000	4
ORL	400	1024	40

6.2 Experimental Setup

Similar with [Yu *et al.*, 2008], we conduct the experiments in the following way. For each data set, we randomly select 50% of the data samples to construct a candidate set, from which

each of the comparing active learning methods is adopted to select a given number $m(= 10, 20, 30, ..., 100)$ of samples. We train a classifier on the selected samples and their labels, and then predict the labels of the remaining 50% data samples. We use the classification accuracy on these 50% data samples to measure the performance of these comparing active learning methods. To avoid the bias of classifier, we choose two classifiers, i.e., SVM with linear kernel and K Nearest Neighbour (KNN) classifier, to predict the class labels. All the experiments are repeated 10 times and the average prediction accuracy is reported.

To evaluate the effectiveness of the proposed Diversified CTED (DCTED for short), we compare DCTED with the following closely related methods, including distribution matching via Maximum Mean Discrepancy (MMD) [Chattopadhyay *et al.*, 2012], Convex Transductive Experimental Design (CTED) [Yu *et al.*, 2008], Active Learning via Neighbourhood Reconstruction (ALNR) [Hu *et al.*, 2013], and Accelerated Robust Subset Selection (ARSS) [Zhu and Fan, 2015]. We also compare DCTED with the method corresponding to Eq. (3). This method leverages a pre-defined and fixed similarity matrix. We denote this method as DCTED$_f$. For DCTED$_f$, we use gaussian function and linear function to define the pairwise similarities among samples and report the best result DCTED$_f$ can achieve. The results of random selection are also reported. To fairly compare the above algorithms, we tune the parameters for all these methods from a large range of $\{10^{-3}, 10^{-2}, ..., 10^{3}\}$. We will give the sensitive analysis of the parameters of DCTED later in this section.

6.3 Experimental Results

We first show the comparisons between the original CTED method and Diversified CTED to investigate whether encouraging diversity can improve the performance of CTED. The experimental results are shown in Tables 2, 3, 4, 5 and 6. The first column of these tables indicate the number of selected samples and C, D$_f$ and D indicate CTED, DCTED$_f$ and DCTED. Each table contains the results based on SVM and KNN classifier. In the reported results, the best one and those having no significant difference (i.e., $p > 0.05$ according to t-test) with the best one are marked in bold. We have the following observations: 1) SVM achieves better performance than KNN classifier in most cases. 2) By encouraging diversity, both DCTED$_f$ and DCTED significantly outperform CTED in most cases. Let us take the USPS data set as an example. When the number of selected samples is 10, DCTED$_f$ and DCTED achieve the accuracy of 0.747 and 0.785, which improve the original CTED method by 29.2% and 35.81%. We also observe that, the original CTED method needs 40 samples to achieve 0.76 while DCTED method only needs 10 samples to achieve 0.785. This indicates that, the set of selected samples by DCTED is more informative since highly similar samples are excluded. 3) DCTED outperforms DCTED$_f$ in most cases. This verifies the effectiveness of DCTED, which adaptively updates the similarity matrix defined on a new data representation. DCTED is more preferable in practice since it does not need to predefine the similarities among samples.

Now we show the comparisons with the state-of-the-art ac-

Table 2: Results on USPS

	SVM			KNN		
m	C	D$_f$	D	C	D$_f$	D
10	0.578	0.747	**0.785**	0.595	0.694	**0.754**
20	0.661	**0.825**	**0.836**	0.658	0.776	**0.806**
30	0.705	**0.849**	**0.859**	0.716	**0.811**	**0.833**
40	0.760	0.864	**0.873**	0.765	**0.832**	**0.848**
50	0.824	**0.874**	**0.884**	0.815	0.849	**0.862**
60	0.854	0.878	**0.890**	0.834	**0.861**	**0.873**
70	0.858	0.882	**0.894**	0.846	0.868	**0.884**
80	0.864	0.884	**0.898**	0.857	0.872	**0.890**
90	0.870	0.887	**0.901**	0.858	0.875	**0.894**
100	0.872	0.890	**0.902**	0.867	0.879	**0.902**

Table 3: Results on MNIST

	SVM			KNN		
m	C	D$_f$	D	C	D$_f$	D
10	0.269	**0.427**	**0.426**	0.277	0.400	**0.431**
20	0.357	**0.537**	**0.549**	0.363	**0.507**	**0.527**
30	0.476	**0.600**	**0.609**	0.466	0.569	**0.591**
40	0.562	**0.633**	**0.643**	0.553	**0.603**	**0.622**
50	0.626	0.655	**0.683**	0.605	0.627	**0.661**
60	0.661	0.676	**0.711**	0.636	0.644	**0.682**
70	0.699	0.691	**0.727**	0.665	0.659	**0.701**
80	0.720	0.696	**0.740**	0.687	0.671	**0.715**
90	**0.735**	0.704	**0.750**	0.700	0.681	**0.724**
100	**0.752**	0.710	**0.760**	0.712	0.689	**0.736**

Table 4: Results on NewsGroup

	SVM			KNN		
m	C	D$_f$	D	C	D$_f$	D
10	0.581	0.611	**0.626**	0.575	**0.596**	**0.614**
20	0.690	0.684	**0.720**	0.663	0.662	**0.694**
30	0.732	0.742	**0.770**	0.690	0.700	**0.717**
40	0.760	0.776	**0.802**	0.715	0.719	**0.736**
50	0.788	0.793	**0.820**	0.730	0.733	**0.749**
60	0.815	0.808	**0.832**	0.742	0.741	**0.760**
70	**0.832**	**0.830**	**0.844**	0.752	0.754	**0.767**
80	0.840	0.838	**0.856**	0.758	0.762	**0.774**
90	0.846	0.847	**0.861**	0.763	0.766	**0.780**
100	**0.853**	**0.853**	**0.863**	**0.772**	0.769	**0.780**

Table 5: Results on WEBKB

	SVM			KNN		
m	C	D$_f$	D	C	D$_f$	D
10	0.595	**0.626**	**0.634**	0.555	**0.609**	**0.609**
20	0.659	**0.685**	**0.693**	0.601	**0.636**	**0.641**
30	0.703	**0.716**	**0.725**	0.632	**0.650**	**0.651**
40	0.716	0.740	**0.754**	0.641	**0.661**	**0.659**
50	0.722	**0.761**	**0.774**	0.647	**0.666**	**0.668**
60	0.732	**0.776**	**0.784**	0.657	**0.671**	**0.675**
70	0.743	0.782	**0.799**	0.659	**0.673**	**0.679**
80	0.753	0.797	**0.808**	0.665	**0.679**	**0.683**
90	0.758	0.806	**0.821**	0.670	**0.684**	**0.688**
100	0.765	0.815	**0.828**	0.674	**0.686**	**0.691**

Table 6: Results on ORL

	SVM			KNN		
m	C	D_f	D	C	D_f	D
10	0.167	0.185	**0.221**	0.163	0.175	**0.211**
20	0.306	0.313	**0.346**	0.274	0.285	**0.327**
30	0.396	**0.418**	**0.439**	0.360	0.377	**0.411**
40	0.467	0.475	**0.519**	0.430	0.436	**0.479**
50	0.524	0.537	**0.589**	0.481	0.495	**0.544**
60	0.562	0.570	**0.636**	0.517	0.525	**0.586**
70	0.601	0.614	**0.698**	0.555	0.567	**0.628**
80	0.622	0.642	**0.728**	0.582	0.604	**0.668**
90	0.666	0.680	**0.766**	0.616	0.628	**0.705**
100	0.700	0.716	**0.786**	0.641	0.655	**0.728**

Table 7: Results Based on SVM

Methods	USPS	MNIST	NewsG	WEBKB	ORL
Random	0.718	0.565	0.695	0.701	0.497
CTED	0.784	0.586	0.774	0.715	0.501
ALNR	0.813	0.492	0.763	0.712	0.494
ARSS	0.847	0.614	0.765	0.715	0.517
MMD	0.841	0.640	0.755	0.719	0.494
$DCTED_f$	0.858	0.633	0.778	0.750	0.515
DCTED	**0.872**	**0.660**	**0.799**	**0.762**	**0.573**

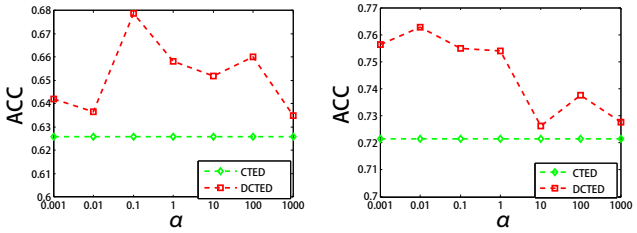
tive learning methods. Due to the space limit, we report the averaged results over different number of selected samples. The results are shown in Tables 7 and 8. From the tables, we see that active learning methods achieve better performance than random selection in most cases. Our method (DCTED) outperforms the other comparing methods on these datasets.

6.4 Parameter Study

Now we study the effects of parameters. Compared with the original CTED method, our DCTED method introduces a new hyper parameter, i.e., α. It is interesting to investigate how this parameter affects the performance of DCTED. To illustrate the effects of α, we fix the number of selected samples as 50. For CTED, we report the best performance it achieves. For our method, i.e., DCTED, we vary the value of α in $[10^{-3}, 10^{-2}, ..., 10^{3}]$ and report the corresponding results. Same as the setting in the previous subsection, the experiments are repeated 10 times with different data partitions and the average results are reported. The results are shown in Figs. 1 and 2. The results in these two figures show that our method (i.e., DCTED) achieves better performance than CTED in a wide range of α. This again verifies that, the ability of CTED can be improved by imposing diversity on sample selection. We also observe that, the performance of DCTED generally becomes better when the value of α increases. In a certain value of α, DCTED achieves the best performance. If we continue to increase the value of α, the performance of DCTED starts to decrease. The reason is that, when α is too large, the diversity regularizer will dominate the objective function and the reconstruction part has little effects on sample selection, leading to select samples which do not have good capacity for data reconstruction.

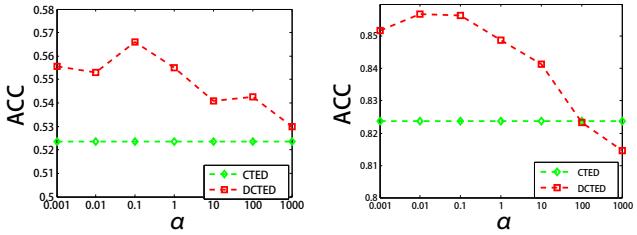
Table 8: Results Based on KNN Classifier

Methods	USPS	MNIST	NewsG	WEBKB	ORL
Random	0.710	0.590	0.638	0.613	0.453
CTED	0.781	0.566	0.716	0.640	0.462
ALNR	0.819	0.475	0.717	0.637	0.456
ARSS	**0.849**	0.590	0.705	0.639	0.475
MMD	0.838	0.617	0.696	0.625	0.450
$DCTED_f$	0.832	0.605	0.720	**0.662**	0.474
DCTED	**0.855**	**0.639**	**0.737**	**0.664**	**0.528**



(a) MNIST (b) WEBKB

Figure 1: Parameter Study on MNIST and WEBKB



(a) ORL (b) USPS

Figure 2: Parameter Study on ORL and USPS

7 Conclusion

In this paper, we proposed to enhance CTED with diversity mechanism by imposing a diversity regularizer over sample selection. Our proposed method, i.e., Diversified CTED, can select informative samples with complementary information and exclude highly similar ones which convey overlapped information. Extensive experimental results demonstrated that, with diversification, Diversified CTED significantly improves CTED and consistently outperforms the state-of-the-art active learning methods.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work is supported in part by China National 973 program 2014CB340301 and NSFC grant 61379043, 61502289.

References

[Balcan *et al.*, 2007] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Learning Theory*, pages 35–50. Springer, 2007.

[Borgwardt *et al.*, 2006] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[Cai and He, 2012] Deng Cai and Xiaofei He. Manifold adaptive experimental design for text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 24(4):707–719, 2012.

[Cai *et al.*, 2006] Deng Cai, Xiaofei He, Jiawei Han, and Hong-Jiang Zhang. Orthogonal laplacianfaces for face recognition. *Image Processing, IEEE Transactions on*, 15(11):3608–3614, 2006.

[Chattopadhyay *et al.*, 2012] Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 741–749. ACM, 2012.

[Chattopadhyay *et al.*, 2013] Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):13, 2013.

[Freund *et al.*, 1997] 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.

[Hu *et al.*, 2013] Yao Hu, Debing Zhang, Zhongming Jin, Deng Cai, and Xiaofei He. Active learning via neighborhood reconstruction. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1415–1421. AAAI Press, 2013.

[Lewis and Catlett, 1994] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[Liu *et al.*, 2010] Jialu Liu, Deng Cai, and Xiaofei He. Gaussian mixture model with local consistency. In *AAAI*, volume 10, pages 512–517. Citeseer, 2010.

[Nguyen and Smeulders, 2004] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79. ACM, 2004.

[Nie *et al.*, 2012] Feiping Nie, Dong Xu, and Xuelong Li. Initialization independent clustering with actively self-training method. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1):17–27, 2012.

[Nie *et al.*, 2013] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Early active learning via robust representation and structured sparsity. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1572–1578. AAAI Press, 2013.

[Settles, 2009] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[Seung *et al.*, 1992] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[Tong and Koller, 2002] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[Wang *et al.*, 2011] Hua Wang, Feiping Nie, Heng Huang, and Fillia Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1553, 2011.

[Wu and Schölkopf, 2006] Mingrui Wu and Bernhard Schölkopf. A local learning approach for clustering. In *Advances in neural information processing systems*, pages 1529–1536, 2006.

[Yang *et al.*, 2014] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2014.

[Yu *et al.*, 2005] Kai Yu, Shipeng Yu, and Volker Tresp. Soft clustering on graphs. In *Advances in neural information processing systems*, pages 1553–1560, 2005.

[Yu *et al.*, 2006] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088. ACM, 2006.

[Yu *et al.*, 2008] Kai Yu, Shenghuo Zhu, Wei Xu, and Yihong Gong. Non-greedy active learning for text categorization using convex ansductive experimental design. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 635–642. ACM, 2008.

[Zhen and Yeung, 2010] Yi Zhen and Dit-Yan Yeung. Sed: supervised experimental design and its application to text classification. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM, 2010.

[Zhu and Fan, 2015] Feiyun Zhu and Bin Fan. 10,000+ times accelerated robust subset selection (arss). In *Proceedings of The Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3217–3223. AAAI Press, 2015.