Discriminative Experimental Design

Yu Zhang and Dit-Yan Yeung

Hong Kong University of Science and Technology
{zhangyu,dyyeung}@cse.ust.hk

Abstract. Since labeling data is often both laborious and costly, the labeled data available in many applications is rather limited. Active learning is a learning approach which actively selects unlabeled data points to label as a way to alleviate the labeled data deficiency problem. In this paper, we extend a previous active learning method called transductive experimental design (TED) by proposing a new unlabeled data selection criterion. Our method, called discriminative experimental design (DED), incorporates both margin-based discriminative information and data distribution information and hence it can be seen as a discriminative extension of TED. We report experiments conducted on some benchmark data sets to demonstrate the effectiveness of DED.

1 Introduction

It is not uncommon that the labeled data available in many machine learning applications is rather limited because the labeling process is both laborious and costly. We refer to this as the labeled data deficiency problem. However, even though labeled data is scarce, abundant unlabeled data may be available in some applications at very low cost. There exist some learning approaches which exploit unlabeled data to boost the generalization performance. One of them is semi-supervised learning [1] which exploits information contained in the unlabeled data such as the geometric structure of the data. Another approach is active learning [2, 3] which expands the labeled data set while keeping the labeling cost low by selecting only the most representative unlabeled data points to label.

Unlike many conventional machine learning methods which wait passively for labeled data to be provided in order to start the learning process, active learning takes a more active approach by selecting unlabeled data points to query some oracle or domain expert. As a result, the expanded labeled data set can help the system learn a better, more accurate model. The typical learning procedure of the active learning approach is depicted in Table 1. Most existing active learning methods, such as support vector machine (SVM) active learning [4–6], select only one data point in each active learning iteration, i.e., the set $\mathcal{S}$ in Table 1 is a singleton set. To select multiple data points, multiple iterations are needed and hence the learning model has to be re-trained multiple times, incurring high computational cost. In recent years, a few active learning methods have been proposed to select multiple data points in each iteration to reduce the total computational cost. Some examples include batch mode active learning [7, 8] and transductive experimental design (TED) [9, 10].

Table 1. Typical Active Learning Procedure

Input: Labeled data set $\mathcal{L}$; Unlabeled data set $\mathcal{U}$ **Output:** Learning model
Step 1: Train a learning model based on $\mathcal{L}$; Step 2: For $t = 1, \ldots, t_{\max}$ 2.1: Select an unlabeled data set $\mathcal{S}$ from $\mathcal{U}$ based on some unlabeled data selection criterion; 2.2: Query an oracle to label $\mathcal{S}$; 2.3: $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{S}, \mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}$; 2.4: Re-train the learning model based on $\mathcal{L}$;

TED has been demonstrated to be an effective method for active learning. However, it can only utilize information about the data distribution. We propose here an extension of TED, called discriminative experimental design (DED), which combines the strengths of both SVM active learning and TED. In particular, the data selection criterion of DED incorporates both margin-based discriminative information and data distribution information. Under the DED framework, we will show that TED can be seen as a special case by treating all data points as equally important. To solve the DED learning problem, we propose a new optimization procedure which exhibits some interesting properties. We will report experiments conducted on some benchmark data sets to demonstrate the effectiveness of DED.

In the next section, we will briefly review active learning and TED. We then present DED in section 3 as a discriminative extension of TED. Section 4 presents our experimental results and then the final section concludes the paper.

2 Active Learning and TED

Among the most important elements of active learning is the unlabeled data selection criterion which has attracted a lot of attention in the machine learning research community. The most commonly used selection criteria include uncertainty sampling [11], query-by-committee [12], representative sampling [13] and Bayesian error reduction [14]. Among these criteria, uncertainty sampling is the most widely studied one. A representative method that uses this criterion is SVM active learning which uses the decision function value as the uncertainty measure for guiding the selection of unlabeled data points. Although SVM active learning performs well in many applications, it does have some limitations. Since it only considers data points lying near the decision boundary of the current classifier, it ignores information about the whole data distribution but such information has been shown to be effective for active learning in representative sampling and TED. Moreover, since labeled data points are often scarce during the early stage of learning, estimation of the margin is not very accurate and hence SVM active learning may select atypical data points or even outliers. Furthermore, since SVM active learning selects only one data point in each iteration, the model has to be re-trained multiple times during the active learning procedure.

TED, which has its origin in experimental design [15] from the statistics community, is used for active learning in [9]. The learning procedure of TED is somewhat different from that of conventional active learning in that it does not assume the existence of labeled data before active learning begins and hence its data selection criterion does not rely on discriminative information provided by the current classifier. By utilizing the data distribution information, TED can choose multiple representative data points in each iteration of the active learning procedure.

It appears that both conventional active learning methods and TED have advantages that are complementary to each other. In the next section, we will propose a new method that combines the strengths of conventional active learning and TED. In particular, the new data selection criterion will incorporate both margin-based discriminative information and data distribution information.

3 Discriminative Experimental Design

Suppose we are given a training set $\mathcal{D}$ which contains both labeled and unlabeled data. The labeled part of $\mathcal{D}$ consists of l labeled data points $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, l$, where $\mathbf{x}_i \in \mathbb{R}^d$ and its corresponding class label $y_i \in \{-1, 1\}$. The unlabeled part of $\mathcal{D}$ consists of u unlabeled data points $\mathbf{x}_j \in \mathbb{R}^d$, $j = l+1, \ldots, l+u$. Usually $l \ll u$ because labeling data is laborious and costly. We assume that the data points are centered and the classification function is defined as $f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$, where $\boldsymbol{\phi}(\cdot)$ denotes the feature map corresponding to some kernel function $k(\cdot, \cdot)$. In general, $\boldsymbol{\phi}(\cdot)$ may have no explicit form but is only defined implicitly.

3.1 Objective Function

As discussed in the previous section, the goal of this paper is to exploit the advantages of both SVM active learning and TED in defining DED. This property should be reflected by the objective function of DED which is what we will look at in this subsection.

The learning setting of DED is similar to that of conventional active learning as depicted in Table 1, which has both labeled and unlabeled data before active learning begins. We hope to define a better data selection criterion which, on one hand, incorporates discriminative information from the labeled data and, on the other hand, incorporates data distribution information from the unlabeled data.

Let us first review the least squares SVM [16] which is used, though in different ways, by both TED and DED. The optimization problem can be stated as follows:

$$\min_{\mathbf{w}} \sum_{i=1}^{l} (\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\|_2^2, \tag{1}$$

where $\lambda > 0$ is the regularization parameter and $\|\cdot\|_2$ denotes the 2-norm for vectors. Since $y_i \in \{-1, 1\}$, (1) is equivalent to the following problem:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \sum_{i=1}^{l} (1 - y_i \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|_2^2. \tag{2}$$

We use the objective function $J(\mathbf{w})$ in (2) to learn the model parameters $\mathbf{w}$ for the classifier. Here we use the squared loss $L(s,t) = (1-st)^2$ which is similar to the squared hinge loss $L'(s,t) = \max(0, 1-st)^2$ used for SVM [17]. Similar to the squared hinge loss, the squared loss used here enforces the prediction of the classifier and the ground truth to have the same sign and that there is a large margin between the positive and negative classes. Moreover, it is equivalent to the conventional squared loss $L(s,t) = (s-t)^2$ for binary classification problems and it is a convex loss. The function score for a test data point is defined as:

$$y = \frac{1}{\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})}, \tag{3}$$

and the final classification decision is based on the sign of the function score. When the denominator is very close to 0, we can add a small value to it to make it numerically more stable.

Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbb{R}^{d \times n}$ denote the matrix for the unlabeled data currently available and the matrix $\mathbf{X} \in \mathbb{R}^{d \times t}$ denote the selected subset of unlabeled data for the oracle to label. So when no action has been taken, n is just equal to u which is the number of unlabeled data points to start with. On the other hand, while conventional active learning methods assume that $l > 0$, TED assumes that $l = 0$ and hence it is not designed to make use of discriminative information.

Problems (1) and (2) are equivalent as far as binary classification problems are concerned. From the derivation of TED, the covariance matrix of the estimation error of $\mathbf{w} - \mathbf{w}^\star$, where $\mathbf{w}^\star$ is the ground truth of $\mathbf{w}$, is proportional to the inverted Hessian matrix of $J(\mathbf{w})$:

$$\begin{aligned}\operatorname{cov}(\mathbf{w} - \mathbf{w}^\star) \propto \mathbf{C}_\mathbf{w} &= \Big(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T}\Big)^{-1} \\ &= \big(\boldsymbol{\phi}(\mathbf{X}) \mathbf{Y}_\mathbf{X}^2 \boldsymbol{\phi}(\mathbf{X})^T + \lambda \mathbf{I}_{d'}\big)^{-1},\end{aligned}$$

where $\mathbf{Y}_\mathbf{X}$ denotes a diagonal matrix whose diagonal elements are the function scores of the corresponding data points, $\mathbf{I}_d$ denotes the $d \times d$ identity matrix, $\boldsymbol{\phi}(\mathbf{X})$ denotes the data matrix of $\mathbf{X}$ after applying the feature map, and d' is the dimensionality of the data points after feature mapping. Then the predictive error on the whole unlabeled data set $\mathbf{V}$ has its covariance matrix proportional to $\mathbf{C}_\mathbf{f}$:

$$\begin{aligned}\mathbf{C}_\mathbf{f} =& \mathbf{Y}_\mathbf{V} \boldsymbol{\phi}(\mathbf{V})^T \mathbf{C}_\mathbf{w} \boldsymbol{\phi}(\mathbf{V}) \mathbf{Y}_\mathbf{V} \\ =& \mathbf{Y}_\mathbf{V} \boldsymbol{\phi}(\mathbf{V})^T \Big(\boldsymbol{\phi}(\mathbf{X}) \mathbf{Y}_\mathbf{X}^2 \boldsymbol{\phi}(\mathbf{X})^T + \lambda \mathbf{I}_{d'}\Big)^{-1} \boldsymbol{\phi}(\mathbf{V}) \mathbf{Y}_\mathbf{V} \\ =& -\frac{1}{\lambda} \mathbf{Y}_\mathbf{V} \boldsymbol{\phi}(\mathbf{V})^T \boldsymbol{\phi}(\mathbf{X}) \mathbf{Y}_\mathbf{X} (\lambda \mathbf{I}_t + \mathbf{Y}_\mathbf{X} \boldsymbol{\phi}(\mathbf{X})^T \boldsymbol{\phi}(\mathbf{X}) \mathbf{Y}_\mathbf{X})^{-1} \mathbf{Y}_\mathbf{X} \boldsymbol{\phi}(\mathbf{X})^T \boldsymbol{\phi}(\mathbf{V}) \mathbf{Y}_\mathbf{V} \\ =& -\frac{1}{\lambda} \mathbf{Y}_\mathbf{V} \mathbf{K}_{\mathbf{VX}} \mathbf{Y}_\mathbf{X} (\lambda \mathbf{I}_t + \mathbf{Y}_\mathbf{X} \mathbf{K}_\mathbf{X} \mathbf{Y}_\mathbf{X})^{-1} \mathbf{Y}_\mathbf{X} \mathbf{K}_{\mathbf{XV}} \mathbf{Y}_\mathbf{V} + \frac{1}{\lambda} \mathbf{Y}_\mathbf{V} \mathbf{K}_\mathbf{V} \mathbf{Y}_\mathbf{V},\end{aligned}$$

where $\mathbf{Y}_\mathbf{V}$ is a diagonal matrix recording the function scores of the data points in $\mathbf{V}$, $\mathbf{K}_\mathbf{X}$ denotes the kernel matrix on $\mathbf{X}$, $\mathbf{K}_\mathbf{V}$ denotes the kernel matrix on $\mathbf{V}$, $\mathbf{K}_{\mathbf{VX}}$ denotes the kernel matrix between $\mathbf{V}$ and $\mathbf{X}$, and $\mathbf{K}_{\mathbf{XV}} = \mathbf{K}_{\mathbf{VX}}^T$. The last equality above holds as a result of the Woodbury identity.

We minimize the predictive variance by using the A-optimal design [15], which minimizes $\mathrm{tr}(\mathbf{C_f})$, the trace of $\mathbf{C_f}$, by treating $\mathrm{tr}(\mathbf{C_f})$ as a surrogate of the predictive variance. Since λ and $\mathbf{Y_V K}_V \mathbf{Y_V}$ are constants, we define the optimization problem for DED as follows.

Definition 1. *Discriminative Experimental Design:*

$$\begin{aligned} \max_{\mathbf{X},\mathbf{Y_X}} \ & \mathrm{tr}\Big[\mathbf{Y_V K_{VX} Y_X}(\lambda \mathbf{I}_t + \mathbf{Y_X K_X Y_X})^{-1}\mathbf{Y_X K_{XV} Y_V}\Big] \\ \textit{s.t.} \ & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = t, \mathbf{Y_X} \subset \mathbf{Y_V}. \end{aligned} \tag{4}$$

Here $\mathbf{X} \subset \mathbf{V}$ means the set of the columns in $\mathbf{X}$ is a subset of that in $\mathbf{V}$ and $|\mathbf{X}|$ denotes the number of data points in $\mathbf{X}$ which is just the number of columns in $\mathbf{X}$. Moreover, for diagonal matrices $\mathbf{Y_X}$ and $\mathbf{Y_V}$, $\mathbf{Y_X} \subset \mathbf{Y_V}$ means the set of the diagonal elements in $\mathbf{Y_X}$ is a subset of that in $\mathbf{Y_V}$.

Before we discuss how to solve problem (4) in the next subsection, let us first examine the relationship between DED and TED. We consider the linear case where

$$\mathbf{K_{VX}} = \mathbf{V}^T\mathbf{X}, \ \mathbf{K_X} = \mathbf{X}^T\mathbf{X}, \ \mathbf{K_{XV}} = \mathbf{X}^T\mathbf{V}.$$

Then the optimization problem for linear DED is

$$\begin{aligned} \max_{\mathbf{X},\mathbf{Y_X}} \ & \mathrm{tr}\Big[\mathbf{Y_V V}^T\mathbf{X Y_X}(\lambda \mathbf{I}_t + \mathbf{Y_X X}^T\mathbf{X Y_X})^{-1}\mathbf{Y_X X}^T\mathbf{V Y_V}\Big] \\ \text{s.t.} \ & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = t, \mathbf{Y_X} \subset \mathbf{Y_V}. \end{aligned}$$

If we define $\tilde{\mathbf{X}} = \mathbf{X Y_X}$ and $\tilde{\mathbf{V}} = \mathbf{V Y_V}$, then the optimization problem for linear DED becomes

$$\begin{aligned} \max_{\tilde{\mathbf{X}}} \ & \mathrm{tr}\Big[\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\lambda \mathbf{I}_t + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\Big] \\ \text{s.t.} \ & \tilde{\mathbf{X}} \subset \tilde{\mathbf{V}}, |\tilde{\mathbf{X}}| = t, \end{aligned}$$

which is exactly the same as TED. So TED can be seen as a special case of DED with $\mathbf{Y_V} = \mathbf{I}_n$. Alternatively, we may regard DED as a weighted version of TED where the weights are related to the function scores of the data points. It is easy to see that both $\mathrm{tr}(\mathbf{C_f})$ and the objective function value of problem (4) do not depend on the signs of the function scores for all data points. So from the definition of function score given in (3), if a data point lies close to the decision boundary, its weight will be much larger than that of another point far from the boundary. A point which lies right at the decision boundary will have the largest weight. This is in line with the design criteria behind SVM active learning and batch mode active learning which use the decision function value as the uncertainty sampling criterion.

Similar to TED, linear DED also has a regularized least squares regression interpretation.

Theorem 1. *Linear Discriminative Experimental Design is equivalent to*

$$\begin{aligned} \min_{\mathbf{X},\mathbf{Y_X},\mathbf{A}} \ & \sum_{i=1}^{n}\Big[\|y_i\mathbf{v}_i - \mathbf{X Y_X a}_i\|_2^2 + \lambda\|\mathbf{a}_i\|_2^2\Big] \\ \textit{s.t.} \ & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = t, \mathbf{Y_X} \subset \mathbf{Y_V} \\ & \mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n) \in \mathbb{R}^{t\times n}. \end{aligned}$$

The proof is similar to that of TED and hence we omit it here. From Theorem 1, we can find that, similar to TED, DED works by selecting representative data points after weighting them with the function scores. Thus DED can utilize both discriminative information and data distribution information to select the most informative data points to label.

3.2 Optimization Procedure

Even though the objective function of DED is similar to that of TED and so, in principle, we may use an optimization method similar to that in [9] to solve problem (4), here we choose to use a different optimization method which gives more insight into the nature of DED and TED.

Let $\boldsymbol{\phi}(\mathbf{X})$ and $\boldsymbol{\phi}(\mathbf{V})$ denote the data matrices after applying the feature map to each data point in $\mathbf{X}$ and $\mathbf{V}$, respectively. From these we get

$$\mathbf{K}_{XV} = \boldsymbol{\phi}(\mathbf{X})^T\boldsymbol{\phi}(\mathbf{V}),\ \mathbf{K}_X = \boldsymbol{\phi}(\mathbf{X})^T\boldsymbol{\phi}(\mathbf{X}),\ \mathbf{K}_V = \boldsymbol{\phi}(\mathbf{V})^T\boldsymbol{\phi}(\mathbf{V}).$$

Since $\boldsymbol{\phi}(\mathbf{X})$ and $\mathbf{Y_X}$ are submatrices of $\boldsymbol{\phi}(\mathbf{V})$ and $\mathbf{Y_V}$ respectively, we can define a selection indicator matrix $\mathbf{S} \in \{0,1\}^{n\times t}$ such that $\boldsymbol{\phi}(\mathbf{X})\mathbf{Y_X} = \boldsymbol{\phi}(\mathbf{V})\mathbf{Y_V}\mathbf{S}$. Because each column of $\boldsymbol{\phi}(\mathbf{X})\mathbf{Y_X}$ is from the column of $\boldsymbol{\phi}(\mathbf{V})\mathbf{Y_V}$, the (i,j)th element s_{ij} of $\mathbf{S}$ can be computed as

$$s_{ij} = \begin{cases} 1 & \text{if } (\boldsymbol{\phi}(\mathbf{X})\mathbf{Y_X})_{,j} \text{ is from } (\boldsymbol{\phi}(\mathbf{V})\mathbf{Y_V})_{,i} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{M}_{,i}$ denotes the ith column of matrix $\mathbf{M}$. Since we need to select t data points from $\mathbf{V}$, one and only one element in each column of $\mathbf{S}$ is equal to 1. Moreover, since we want to select t distinct data points from $\mathbf{V}$, at most one element in each row of $\mathbf{S}$ is equal to 1. In other words, $\mathbf{S}$ consists of t distinct columns of the $n \times n$ identity matrix. So the columns of $\mathbf{S}$ consist of an orthogonal basis such that $\mathbf{S}^T\mathbf{S} = \mathbf{I}_t$. The constraint set for $\mathbf{S}$ can be defined as

$$C_S = \left\{\mathbf{S} \,|\, \mathbf{S} \in \{0,1\}^{n\times t}, \mathbf{S}^T\mathbf{1}_n = \mathbf{1}_t, \mathbf{S}\mathbf{1}_t \leq \mathbf{1}_n\right\},$$

or equivalently

$$C_S = \left\{\mathbf{S} \,|\, \mathbf{S} \in \{0,1\}^{n\times t}, \mathbf{S}^T\mathbf{S} = \mathbf{I}_t\right\},$$

where $\mathbf{1}_m$ denotes an $m \times 1$ vector of all ones and $\leq$ refers to the elementwise comparison between two vectors. Then we can get

$$\begin{aligned} \mathbf{Y_V}\mathbf{K_{VX}}\mathbf{Y_X} &= \mathbf{Y_V}\boldsymbol{\phi}(\mathbf{V})^T\boldsymbol{\phi}(\mathbf{X})\mathbf{Y_X} \\ &= \mathbf{Y_V}\boldsymbol{\phi}(\mathbf{V})^T\boldsymbol{\phi}(\mathbf{V})\mathbf{Y_V}\mathbf{S} \\ &= \mathbf{Y_V}\mathbf{K_V}\mathbf{Y_V}\mathbf{S} \\ \mathbf{Y_X}\mathbf{K_X}\mathbf{Y_X} &= \mathbf{Y_X}\boldsymbol{\phi}(\mathbf{X})^T\boldsymbol{\phi}(\mathbf{X})\mathbf{Y_X} \\ &= \mathbf{S}^T\mathbf{Y_V}\boldsymbol{\phi}(\mathbf{V})^T\boldsymbol{\phi}(\mathbf{V})\mathbf{Y_V}\mathbf{S} \\ &= \mathbf{S}^T\mathbf{Y_V}\mathbf{K_V}\mathbf{Y_V}\mathbf{S}. \end{aligned}$$

Thus the objective function in (4) becomes

$$\begin{aligned} \max_{\mathbf{S}} \quad & \mathrm{tr}\Big[(\lambda \mathbf{I}_t + \mathbf{S}^T \tilde{\mathbf{K}}_{\mathbf{V}} \mathbf{S})^{-1} \mathbf{S}^T \tilde{\mathbf{K}}_{\mathbf{V}}^2 \mathbf{S}\Big] \\ \text{s.t.} \quad & \mathbf{S} \in C_S, \end{aligned} \tag{5}$$

where $\tilde{\mathbf{K}}_{\mathbf{V}} = \mathbf{Y}_{\mathbf{V}} \mathbf{K}_{\mathbf{V}} \mathbf{Y}_{\mathbf{V}}$. By imposing a constraint on $\mathbf{S}$ such that $\mathbf{S}^T \mathbf{S} = \mathbf{I}_t$, (5) can be rewritten as

$$\begin{aligned} \max_{\mathbf{S}} \quad & \mathrm{tr}\Big[\Big(\mathbf{S}^T (\lambda \mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}}) \mathbf{S}\Big)^{-1} \mathbf{S}^T \tilde{\mathbf{K}}_{\mathbf{V}}^2 \mathbf{S}\Big] \\ \text{s.t.} \quad & \mathbf{S} \in C_S. \end{aligned} \tag{6}$$

The formulation of the objective function in (6) is identical to that of linear discriminant analysis (LDA) [18], so the optimal solution can be obtained by solving a generalized eigenvalue problem if there exist no constraints on $\mathbf{S}$.

Here we use a projection method to solve problem (6). That is, we first find the optimal solution of problem (6) while ignoring the constraints and then project the optimal solution found to the constraint set C_S.

We first solve the problem

$$\max_{\mathbf{S}} \; f(\mathbf{S}) = \Big[\Big(\mathbf{S}^T (\lambda \mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}}) \mathbf{S}\Big)^{-1} \mathbf{S}^T \tilde{\mathbf{K}}_{\mathbf{V}}^2 \mathbf{S}\Big]. \tag{7}$$

According to the analysis in [18], the optimal solution $\mathbf{S}^\star$ consists of the top t eigenvectors of $(\lambda \mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}})^{-1} \tilde{\mathbf{K}}_{\mathbf{V}}^2$. Let $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_n]$ and $\boldsymbol{\Pi} = \mathrm{diag}(\pi_1, \ldots, \pi_n)$ denote the eigenvectors and eigenvalues, respectively, of the matrix $\tilde{\mathbf{K}}_{\mathbf{V}}$ where $\pi_1 \geq \ldots \geq \pi_n$. Then, by using the fact that $\mathbf{Q}$ is an $n \times n$ orthogonal matrix because $\mathbf{Q}$ is the eigenvector matrix of a symmetric matrix $\tilde{\mathbf{K}}_{\mathbf{V}}$, we can get

$$\begin{aligned} (\lambda \mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}})^{-1} \tilde{\mathbf{K}}_{\mathbf{V}}^2 &= (\lambda \mathbf{I}_n + \mathbf{Q}^T \boldsymbol{\Pi} \mathbf{Q})^{-1} \mathbf{Q}^T \boldsymbol{\Pi}^2 \mathbf{Q} \\ &= (\lambda \mathbf{Q}^T \mathbf{Q} + \mathbf{Q}^T \boldsymbol{\Pi} \mathbf{Q})^{-1} \mathbf{Q}^T \boldsymbol{\Pi}^2 \mathbf{Q} \\ &= \mathbf{Q}^T (\lambda \mathbf{I}_n + \boldsymbol{\Pi})^{-1} \mathbf{Q} \mathbf{Q}^T \boldsymbol{\Pi}^2 \mathbf{Q} \\ &= \mathbf{Q}^T (\lambda \mathbf{I}_n + \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^2 \mathbf{Q}. \end{aligned}$$

So $\mathbf{q}_i$ is an eigenvector of $(\lambda \mathbf{I}_n + \tilde{\mathbf{K}}_{\mathbf{V}})^{-1} \tilde{\mathbf{K}}_{\mathbf{V}}^2$ with the corresponding eigenvalue as

$$\pi_i' = h(\pi_i) = \pi_i^2 / (\lambda + \pi_i).$$

Then $\mathbf{S}^\star$ consists of t eigenvectors with the t largest eigenvalues in $\{\pi_i'\}$. We find that $h(x)$ is monotonically increasing for $x \geq 0$ since $h'(x) = \frac{x^2 + 2\lambda x}{(x+\lambda)^2} \geq 0$ given $\lambda > 0$, and $h(x)$ is strictly increasing when $x > 0$. So we have $\pi_i' > \pi_j'$ when $\pi_i > \pi_j \geq 0$ and $\mathbf{S}^\star$ consists of the top t eigenvectors of $\tilde{\mathbf{K}}_{\mathbf{V}}$.

We next project $\mathbf{S}^\star$ to the set C_S. Note that $\mathbf{S}^\star$ is not a unique optimal solution of problem (7) since for any orthogonal matrix $\mathbf{P} \in \mathbb{R}^{t \times t}$ we have $f(\mathbf{S}^\star \mathbf{P}) = f(\mathbf{S}^\star)$. So we define the objective function for the projection as

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{Q}} \quad & \|\mathbf{S}^\star \mathbf{P} - \mathbf{Q}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Q} \in C_S, \; \mathbf{P} \mathbf{P}^T = \mathbf{I}_t, \end{aligned} \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. We simplify the objective function as

$$\begin{aligned}\|\mathbf{S}^\star\mathbf{P}-\mathbf{Q}\|_F^2 &= \mathrm{tr}\Big((\mathbf{S}^\star\mathbf{P}-\mathbf{Q})^T(\mathbf{S}^\star\mathbf{P}-\mathbf{Q})\Big)\\ &= \mathrm{tr}(\mathbf{Q}^T\mathbf{Q})+\mathrm{tr}(\mathbf{P}^T(\mathbf{S}^\star)^T\mathbf{S}^\star\mathbf{P})-2\mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})\\ &= 2t-2\,\mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P}).\end{aligned}$$

Note that the last equality holds because $\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_t$, $(\mathbf{S}^\star)^T\mathbf{S}^\star=\mathbf{I}_t$ and $\mathbf{P}^T\mathbf{P}=\mathbf{I}_t$. So minimizing $\|\mathbf{S}^\star\mathbf{P}-\mathbf{Q}\|_F^2$ is equivalent to maximizing $\mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})$ and so problem (8) is equivalent to the following problem

$$\begin{aligned}\max_{\mathbf{P},\mathbf{Q}}\ & \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})\\ \text{s.t.}\ & \mathbf{Q}\in C_S,\ \mathbf{P}\mathbf{P}^T=\mathbf{I}_t.\end{aligned}\tag{9}$$

However, problem (9) is not convex. Here we use an alternating method to solve it. Specifically, we first find the optimal solution with respect to $\mathbf{Q}$ when $\mathbf{P}$ is fixed and then find the optimal solution with respect to $\mathbf{P}$ when $\mathbf{Q}$ is fixed.

When $\mathbf{P}$ is fixed, the optimization problem with respect to $\mathbf{Q}$ is

$$\begin{aligned}\max_{\mathbf{Q}}\ & \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})\\ \text{s.t.}\ & \mathbf{Q}\in\{0,1\}^{n\times t},\mathbf{Q}^T\mathbf{1}_n=\mathbf{1}_t,\mathbf{Q}\mathbf{1}_t\le\mathbf{1}_n.\end{aligned}\tag{10}$$

This problem is to find the t largest elements in $\mathbf{S}^\star\mathbf{P}$ where no two elements can be in the same column or the same row. This is an integer programming problem with no efficient solution. Based on our observation that the largest elements of different columns in $\mathbf{S}^\star$ usually lie in different rows, we propose a greedy algorithm for the problem: we first find the largest element in $\mathbf{S}^\star\mathbf{P}$ (if there exist multiple elements that are the largest, we can choose any one of them) and mark its row and column; then from the unmarked columns and rows we find the largest one and also mark it; this procedure is repeated until we find t elements.

When $\mathbf{Q}$ is fixed, the optimization problem with respect to $\mathbf{P}$ is

$$\begin{aligned}\max_{\mathbf{P}}\ & \mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})\\ \text{s.t.}\ & \mathbf{P}\mathbf{P}^T=\mathbf{I}_t.\end{aligned}\tag{11}$$

We define a Lagrangian using a symmetric matrix multiplier $\boldsymbol{\Lambda}$ as

$$L(\mathbf{P},\boldsymbol{\Lambda})=\mathrm{tr}(\mathbf{Q}^T\mathbf{S}^\star\mathbf{P})-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}(\mathbf{P}\mathbf{P}^T-\mathbf{I}_t)).$$

Then the optimal solution $(\mathbf{P}^\star,\boldsymbol{\Lambda}^\star)$ satisfies

$$\frac{\partial L}{\partial\mathbf{P}}=(\mathbf{S}^\star)^T\mathbf{Q}-\boldsymbol{\Lambda}^\star\mathbf{P}^\star=0$$

which leads to $\boldsymbol{\Lambda}^\star=(\mathbf{S}^\star)^T\mathbf{Q}(\mathbf{P}^\star)^T$ by right-multiplying $(\mathbf{P}^\star)^T$. Utilizing the fact that $\mathbf{P}^\star$ is a $t\times t$ orthogonal matrix, we get

$$\boldsymbol{\Lambda}^\star(\boldsymbol{\Lambda}^\star)^T=(\mathbf{S}^\star)^T\mathbf{Q}\mathbf{Q}^T\mathbf{S}^\star.$$

Let $(\mathbf{S}^\star)^T\mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{R}^T$ be the singular value decomposition (SVD) where $\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{R} \in \mathbb{R}^{t\times t}$. So

$$\boldsymbol{\Lambda}^\star(\boldsymbol{\Lambda}^\star)^T = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T$$

and

$$\boldsymbol{\Lambda}^\star = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$$

since $\boldsymbol{\Lambda}^\star$ is symmetric. Then we can get the optimal $\mathbf{P}^\star$ as

$$\mathbf{P}^\star = (\boldsymbol{\Lambda}^\star)^{-1}(\mathbf{S}^\star)^T\mathbf{Q} = \mathbf{U}\mathbf{R}^T.$$

The main computational cost includes computing the eigenvectors of $\tilde{\mathbf{K}}$ corresponding to the largest t eigenvalues only one time which costs $O(n^2t)$ and SVD for $(\mathbf{S}^\star)^T\mathbf{Q}$ which costs $O(t^3)$. So the computational cost of our method is $O(n^2t)$, which is more efficient than that of [9] with $O(n^3)$ complexity. Moreover, our method provides a better characterization of the nature of DED. The regularization parameter λ has significant effect on the optimization procedure of TED in [9] but DED seems to be insensitive to it. This property is desirable because DED is robust against λ.

4 Experiments

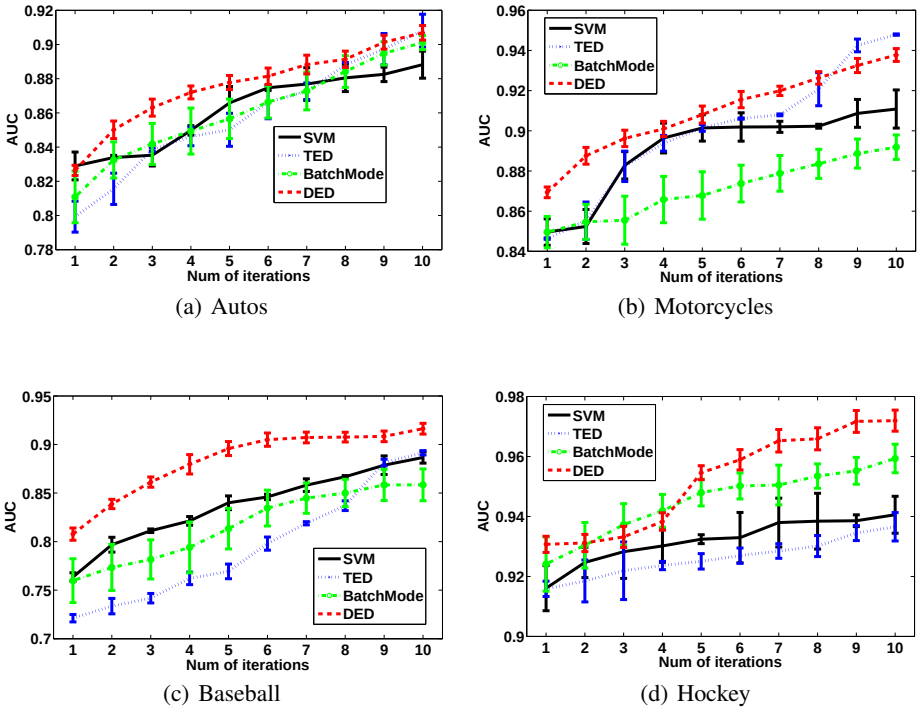
In this section, we study DED empirically and compare its performance with several active learning methods, which include TED, SVM active learning and batch mode active learning [19].

We conduct experiments on two public benchmark data sets. The first one is a subset of the Newsgroups corpus [20], which consists of 3970 documents with TFIDF features of 8014 dimensions. Each document belongs to exactly one of four categories: autos, motorcycles, baseball and hockey. The other one is the Reuters data set, which is a subset of the RCV1-v2 data set [21]. Each document in the Reuters data set belongs to at least one of four categories: CCAT, ECAT, GCAT and MCAT.

In the experiments, we simply treat the multi-class/label classification problem as a set of binary classification problems by using the one-versus-all scheme, i.e., documents from the target category are labeled as positive examples and those from the other categories are labeled as negative examples. We use area under the ROC curve (AUC) as the performance measure to measure the overall classification performance, because in our setting, each binary classification task is unbalanced (only about 25% of the documents in the Newsgroups data set and about 30% of the documents in the Reuters data set are positive).

In our experiments, t is set to 5 and all the regularization parameters in DED, TED and SVM active learning are set to 0.01. We initially have five labeled data points for each class before active learning starts.

We first test our method on the Newsgroups data set. The AUC values over the four binary classification tasks are reported in Figure 1(a) to 1(d). From the results, we can see that on Autos, Motorcycles and Baseball, DED outperforms the other methods in the early stage. This observation validates the contribution of data distribution information. When the labeled data is scarce, data distribution information may be more important



(a) Autos (b) Motorcycles

(c) Baseball (d) Hockey

Fig. 1. Learning curves for four binary classification tasks on Newsgroups data

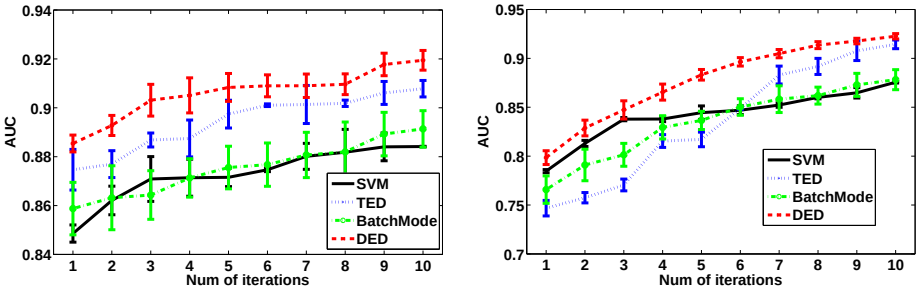
than discriminative information since the estimated decision boundary in this stage is not very accurate.

We now compare the four methods on the Reuters data set. The AUC values over the four tasks are reported in Figure 2(a) to 2(d). For the four categories, DED consistently outperforms the second best by a large margin. This observation validates the contribution of discriminative information to experimental design.

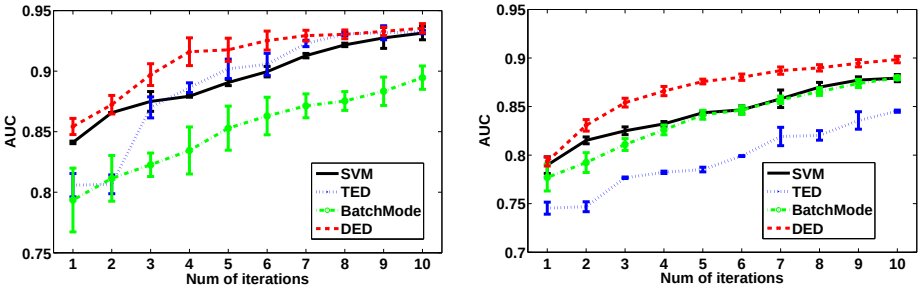
Moreover, to see the effect of the optimization method proposed in section 3.2, we compare the performance of DED when using our proposed optimization method and the one proposed in [9]. The AUC values averaged over four binary classification tasks of the two data sets are reported in Fig. 3(a) and Fig. 3(b). We can see that the performance of our proposed method is better than the one proposed in [9].

5 Conclusion

We have proposed in this paper a novel active learning method which integrates margin-based discriminative information and data distribution information to define the unlabeled data selection criterion. As the next step to extend this work further, we will investigate the integration of active learning and semi-supervised learning to further improve the performance by exploiting unlabeled data.

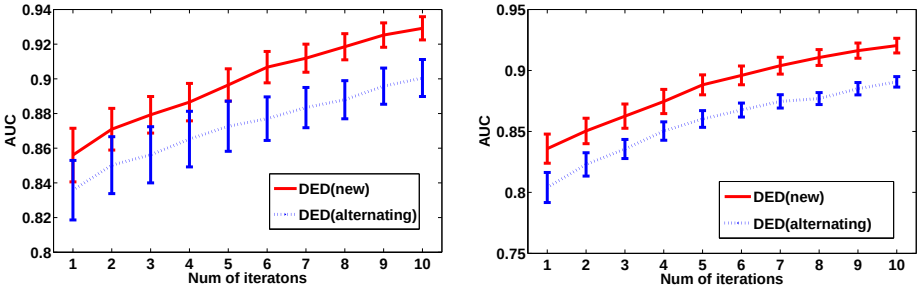


(a) CCAT (b) ECAT



(c) GCAT (d) MCAT

Fig. 2. Learning curves for four binary classification tasks on Reuters data



(a) Newsgroups data (b) Reuters data

Fig. 3. Comparison of two optimization methods for DED on the two data sets. DED(new) uses the optimization procedure proposed in our paper and DED(alternating) utilizes the alternating optimization method proposed in [9].

Acknowledgment

This research has been supported by General Research Fund 622209 from the Research Grants Council of Hong Kong.

References

1. Chapelle, O., Zien, A., Schölkopf, B., eds.: Semi-Supervised Learning. MIT Press, Boston (2006)
2. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning **15** (1994) 201–221
3. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence Research **4** (1996) 129–145
4. Campbell, C., Cristianini, N., Smola, A.J.: Query learning with large margin classifiers. In: Proceedings of the Seventeenth International Conference on Machine Learning, Stanford University, Standord, CA, USA (2000) 111–118
5. Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: Proceedings of the Seventeenth International Conference on Machine Learning, Stanford University, Standord, CA, USA (2000) 839–846
6. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: Proceedings of the Seventeenth International Conference on Machine Learning, Stanford University, Standord, CA, USA (2000) 999–1006
7. Hoi, S.C.H., Jin, R., Zhu, J., Lyu, M.R.: Batch mode active learning and its application to medical image classification. In: Proceedings of the Twenty-Third International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA (2006) 417–424
8. Guo, Y., Schuurmans, D.: Discriminative batch mode active learning. In Platt, J., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems 20. (2007) 593–600
9. Yu, K., Bi, J., Tresp, V.: Active learning via transductive experimental design. In: Proceedings of the Twenty-Third International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA (2006) 1081–1088
10. Sindhwani, V., Melville, P., Lawrence, R.D.: Uncertainty sampling and transductive experimental design for active dual supervision. In: Proceedings of the 26th International Conference on Machine Learning, Montreal, Quebec, Canada (2009) 953–960
11. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Proceedings of the 11th International Conference on Machine Learning, Morgan Kaufmann (1994) 148–156
12. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory, New York, NY, USA (1992) 287–294
13. Nguyen, H.T., Smeulders, A.: Active learning using preclustering. In: Proceedings of the 21st International Conference on Machine learning, Banff, Alberta, Canada (2004)
14. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, USA (2001) 441–448
15. Atkinson, A.C., Donev, A.N., eds.: Optimum Experiment Designs. Oxford University Press, Boston (1992)
16. Gestel, T.V., Suykens, J.A.K., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., Moor, B.D., Vandewalle, J.: Benchmarking least squares support vector machine classifiers. Machine Learning **54**(1) (2004) 5–32
17. Hsieh, C.J., Chang, K.W., Lin, C.J., Keerthi, S.S., Sundararajan, S.: A dual coordinate descent method for large-scale linear SVM. In: Proceedings of the Twenty-Fifth International Conference on Machine Learning, Helsinki, Finland (2008) 408–415
18. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, New York (1991)

19. Hoi, S.C.H., Jin, R., Zhu, J., Lyu, M.R.: Semi-supervised svm batch mode active learning for image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA (2008)
20. Yu, K., Zhu, S., Xu, W., Gong, Y.: 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, Singapore (2008) 635–642
21. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5** (2004) 361–397