Non-greedy Active Learning for Text Categorization using Convex Transductive Experimental Design

Kai Yu, Shenghuo Zhu, Wei Xu, Yihong Gong
NEC Laboratories America, Cupertino, California 95014
{kyu, zsh, wx, ygong}@sv.nec-labs.com

ABSTRACT

In this paper we propose a non-greedy active learning method for text categorization using least-squares support vector machines (LSSVM). Our work is based on transductive experimental design (TED), an active learning formulation that effectively explores the information of unlabeled data. Despite its appealing properties, the optimization problem is however NP-hard and thus—like most of other active learning methods—a greedy sequential strategy to select one data example after another was suggested to find a suboptimum. In this paper we formulate the problem into a continuous optimization problem and prove its convexity, meaning that a set of data examples can be selected with a guarantee of global optimum. We also develop an iterative algorithm to efficiently solve the optimization problem, which turns out to be very easy-to-implement. Our text categorization experiments on two text corpora empirically demonstrated that the new active learning algorithm outperforms the sequential greedy algorithm, and is promising for active text categorization applications.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms

Algorithms, Theory, Performance

Keywords

Active Learning, Convex Optimization, Text Categorization, Transductive Experimental Design

1. INTRODUCTION

There has been a long tradition of research on active learning for text classification. In order to train a classification model that can automatically assign semantic tags on documents, usually human experts need to provide a large set of labeled examples. Active learning reduces the labeling costs by identifying and presenting the most informative examples for experts to label.

Despite of a large body of work done by researchers, most of active learning algorithms are still far from being satisfactory and have apparent shortcomings. Many methods do not explore the information about the distribution of unlabeled data. As another drawback, nearly all the algorithms take greedy strategies to iteratively select one examples after another, which is however suboptimal compared with optimizing a set of selections at a time.

In this paper we propose a non-greedy active learning method for text categorization using least-squares support vector machines (LSSVM). Our work is based on transductive experimental design (TED) [17], an active learning formulation that effectively explores the information of unlabeled data. Despite its appealing properties, the optimization problem is however NP-hard and thus—like most of other active learning methods—a greedy strategy to select one data example after another was suggested to find a suboptimum. In this paper we transform TED problem into an equivalent form and further replace the cardinality constraint by a novel sparsity regularization. The original discrete problem then becomes a continuous optimization problem. A bit surprisingly, the new formulation is *convex*, meaning that a globally optimal set of data examples can be selected. To the best of our knowledge, few attentions have been put to active learning algorithm that can select multiple data examples simultaneously with a global optimality. We describe an efficient learning algorithm that is very easy to implement. Our text categorization experiments on two text corpora empirically demonstrated that the new active learning algorithm is superior in comparisons with competitive methods.

The paper is organized as follows. In Section 2 we briefly review the related work in active learning. In Section 3 we begin by introducing the transductive experimental design and then propose the idea of convex transductive experimental design, where we prove the convexity and describe the algorithm. We also discuss how to control the sparsity of the result. Finally we empirically evaluate the suggested method in Section 4 and conclude this work in Section 5.

2. RELATED WORK

There has been extensive research on the subject of active learning. Existing approaches either select the most uncertain data given previously trained models [5], or choose the most informative data that optimize some expected gain

SIGIR'08, July 20–24, 2008, Singapore.

[3, 10, 2]. The latter typically requires expensive retraining of models when evaluating each candidate. Some other approaches assume generative models and explore the dependency between inputs and outputs [11]. Active learning methods for Gaussian processes [6] has also been suggested.

[16] proposed active learning methods for support vector machines. The method queries points to reduce the size of the version space as much as possible. As the difficulty in measure the version space, they provided three ways of approximating the procedure, Simple Margin, MaxMin Margin and Ratio Margin. The simple margin method, which selects the example closest to the current decision boundary, was also proposed by [13] and has been very popular. However, the method tends to select untypical data points, which may lead to a poor performance.

Active learning is also referred to as *experimental design* in statistics [1]. In order to learn a predictive function from *experiment-measurements* pairs, experimental design selects the most informative experiments to measure, given that conducting an experiment is expensive. [12] proposes an experimental design method based on logistic regression models. The goal is to minimize the variance. As the variance estimation depends on the label value of data point to be labeled, a factitious label has to be added into the training data before evaluating the variance reduction for each candidate data point. This procedure takes a lot of computation power when the size of the candidate set is large. [8] also investigated the active learning problem for logistic regression models. Their method requires non-trivial optimization techniques to solving submodule problems.

Usual experimental design methods aim to reduce the uncertainty of models (or model parameters), transductive experimental design (TED) was proposed in [17] to directly reduce the assessed uncertainty of predictions on given unlabeled data, and thus effectively explore the information of unlabeled data in active learning. The method was applied to active learning for sponsored search [20]. A related approach was applied in image retrieval [7]. However, despite its appealing performance, the problem is essentially NP-hard, and was thus solved by a greedy sequential algorithm that each time selects only one data example.

3. ACTIVE LEARNING WITH CONVEX TED

3.1 The Setting

Suppose that we have a binary classification problem. A classifier is expected to predict the relationship from the feature vector $\mathbf{x}$ of a document to its labels $y \in \{-1, 1\}$ via

$$y = \operatorname{sign}(f(\mathbf{x})) \tag{1}$$

where the function is assumed to be $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ in this paper. We note that a bias term can be incorporated into the form by expanding the weights and input feature vector as $\mathbf{x} \leftarrow [x, 1]$ and $\mathbf{w} \leftarrow [\mathbf{w}, b]$. Based on a set of training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^K$, the so-called least-squares support vector machine [14] (LSSVM) is equivalent to least-square ridge regression, which, in linear case, learns $f(\mathbf{x})$ by estimating $\mathbf{w}$ via

$$\mathbf{w}^* = \min_{\mathbf{w}} \left\{ J(\mathbf{w}; \mathbf{X}_\mathcal{A}) = \sum_{i \in \mathcal{A}} \left(\mathbf{w}^\top \mathbf{x}_i - y_i \right)^2 + \mu \|\mathbf{w}\|^2 \right\} \tag{2}$$

where $\mu > 0$, $\|\cdot\|$ is the vector 2-norm, $\mathbf{X}_\mathcal{A} = \{\mathbf{x}_i\}_{i \in \mathcal{A}}$ is a set of $|\mathcal{A}| = K$ training examples. The method has shown state-of-art text categorization performance [19, 18].

A generic active learning problem aims to choose an optimal training set $\mathbf{X}_\mathcal{A}$, what we call *active set* in this paper, from a set of candidates $\mathbf{X}_\mathcal{C}$, $|\mathcal{C}| = N$, such that some quality measurement of $\mathbf{w}^*$ is maximized.

Throughout this paper, we will somewhat abuse the notation, for example, $\mathbf{X}_\mathcal{A}$ and $\mathbf{X}_\mathcal{C}$ represent sets, but may also be used to denote the matrices $(\mathbf{x}_i)_{i \in \mathcal{A}}$ and $(\mathbf{x}_j)_{j \in \mathcal{C}}$, respectively. Their meanings should be clear given their contexts. If the feature vector $\mathbf{x}$ be D-dimensional, then $\mathbf{X}_\mathcal{A} \in \mathbb{R}^{K \times D}$. $|\mathcal{A}|$ denotes the size of set $\mathcal{A}$, and $|a|$ denotes the absolute value of a if a is a scalar. Moreover, we use $\|\boldsymbol{\beta}\|_0$, $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|$ to represent the ℓ_0-norm, ℓ_1-norm and ℓ_2-norm of vector $\boldsymbol{\beta}$, respectively. We note that $\|\boldsymbol{\beta}\|_0$ is the number of nonzero elements in $\boldsymbol{\beta}$.

3.2 Transductive Experimental Design

Based on the learning method (2), the key idea of TED [17] is to minimize the *average predictive variance* of the learned function $f(\mathbf{x})$ on pre-given data $\mathbf{X}_\mathcal{P}$, $|\mathcal{P}| = M$, which are to be predicted. It is formulated as an optimization problem

$$\min_{\mathcal{A} \subset \mathcal{C}, |\mathcal{A}| = K} \frac{1}{M} \operatorname{trace} \left(\mathbf{X}_\mathcal{P} \mathbf{H}^{-1} \mathbf{X}_\mathcal{P}^\top \right) \tag{3}$$

where $\mathbf{H}$ is the Hessian matrix

$$\mathbf{H} = \frac{\partial J(\mathbf{w}; \mathbf{X}_\mathcal{A})}{\partial \mathbf{w} \partial \mathbf{w}^\top} = \mathbf{X}_\mathcal{A} \mathbf{X}_\mathcal{A}^\top + \mu \mathbf{I} \tag{4}$$

In many applications, like text categorization, it is reasonable to assume a large availability of unlabeled data $\mathbf{X}_\mathcal{P}$, whose distribution should be taken into account to effectively impact the choice of the active set $\mathcal{A}$. Generally the candidates $\mathbf{X}_\mathcal{C}$ can be a subset of $\mathbf{X}_\mathcal{P}$, or a completely different set, or simply $\mathbf{X}_\mathcal{C} = \mathbf{X}_\mathcal{P}$.

Somewhat surprisingly, the predictive variance of $f(\mathbf{x}) = \langle \mathbf{w}^*, \mathbf{x} \rangle$ depends only on the input features of training examples, due to the fact that, in $J(\mathbf{w}; \mathbf{X}_\mathcal{A})$ labels y_i only linearly couple with $\mathbf{w}$ and hence a second order derivative with respect to $\mathbf{w}$ has all the y_i terms canceled out. This independence of y_i removes the complication of the unknown factor of human labeling in active learning process. After some mathematical derivation, the minimization of predictive variance can be formulated as an equivalent optimization problem with a cardinality constraint

$$\begin{aligned} \min_{\mathcal{A}, \boldsymbol{\alpha}_i \in \mathbb{R}^K} \quad & \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_\mathcal{A}^\top \boldsymbol{\alpha}_i\|^2 + \mu \|\boldsymbol{\alpha}_i\|^2 \\ \text{subject to} \quad & |\mathcal{A}| = K, \quad \mathcal{A} \subset \mathcal{C}, \quad \mathbf{x}_i \in \mathbf{X}_\mathcal{P} \end{aligned} \tag{5}$$

which reveals that TED aims to find the optimal common set of K active examples $\mathbf{X}_\mathcal{A}$ to approximate every test data $\mathbf{x}_i \in \mathbf{X}_\mathcal{P}$ by $\hat{\mathbf{x}}_i = \mathbf{X}_\mathcal{A}^\top \boldsymbol{\alpha}_i$, $i \in \mathcal{P}$. The approximation can be seen as the (regularized) projection of $\mathbf{x}_i$ onto the linear subspace spanned by $\mathbf{X}_\mathcal{A}$. Therefore, TED has a geometric interpretation that it tends to find *representative* data $\mathbf{X}_\mathcal{A}$ spanning a linear subspace to retain most of the information of the whole set of test data $\mathbf{X}_\mathcal{P}$. Therefore, given a sufficiently large set $\mathbf{X}_\mathcal{P}$, TED actually explores the information about the distribution of unlabeled data.

Despite of the appealing interpretation, (5) is however an NP-hard problem. Two suboptimal solutions have been suggested so far. The first is a sequential algorithm that solves a problem (5) of size $K = 1$ at each step, to approximate the residuals of the previous step. The algorithm is described as Algorithm 1.

Algorithm 1 Sequential TED

```
Require: candidates $\mathbf{X}_\mathcal{C}$, unlabeled data $\mathbf{X}_\mathcal{P}$, $\mu > 0$, $K$;
initialize $\kappa_{i,j} \leftarrow \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, for $i \in \mathcal{P}, j \in \mathcal{C}$;
repeat
    $j \leftarrow \arg\max_{j \in \mathcal{C}} \sum_{i \in \mathcal{P}} \kappa_{i,j}^2 / (\kappa_{j,j} + \mu)$;
    $\mathcal{A} \leftarrow \mathcal{A} \cup j$;
    $\kappa_{i,i'} \leftarrow \kappa_{i,i'} - \kappa_{i,j}\kappa_{i',j}/(\kappa_{j,j} + \mu)$, for $i \in \mathcal{P}, i' \in \mathcal{C}$;
until $|\mathcal{A}| = K$;
return $\mathbf{X}_\mathcal{A}$;
```

The second approach aims to optimize the set $\mathbf{X}_\mathcal{A}$ simultaneously, which replaces the cardinality constraint $|\mathcal{A}| = K$ by an $\ell 1$-norm regularization

$$\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}_i \in \mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_\mathcal{C}^\top \mathbf{B} \boldsymbol{\alpha}_i\|^2 + \mu \|\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \gamma \|\boldsymbol{\beta}\|_1$$

$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_\mathcal{P}, \quad \mathbf{B} = \operatorname{diag}(\boldsymbol{\beta}), \quad \mathbf{B} \succeq 0. \tag{6}$$

The above formulation introduces a set of variables β_j to control the overall "on" and "off" of each candidate $\mathbf{x}_j$ in terms of being selected or not. The $\ell 1$-norm $\|\boldsymbol{\beta}\|_1$ enforces some elements of $\boldsymbol{\beta}$ to be zero. The optimization is done by alternatively optimizing β_j or $\boldsymbol{\alpha}_j$ while fixing the other. This problem is however non-convex, which means that the results are highly sensitive to the initialization of the algorithm, and can be easily trapped into a poor local minimum.

3.3 A Convex Formulation

In this section we introduce a new formulation of TED, which is convex and hence avoids the risk of being trapped into any local optimum. Unlike most of the other active learning algorithms, the new method aims to select *multiple* examples at a time with a guarantee of global optimum. We will also show that the optimization procedure is actually easy to implement.

Again, we introduce auxiliary variables $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]$ to control the inclusion of examples into the training set. The optimization problem is

$$\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}_i \in \mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_\mathcal{C}^\top \boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{N} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma \|\boldsymbol{\beta}\|_1 \tag{7}$$

$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_\mathcal{P}, \quad \beta_j \geq 0, \quad j = 1, \cdots, N,$$

where $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,N}]^\top$. As suggested by the well-known LASSO method [15], the $\ell 1$-norm $\|\boldsymbol{\beta}\|_1$ will enforce some elements of $\boldsymbol{\beta}$ to be zero. It is not difficult to see, if $\beta_j = 0$, then all $\alpha_{1,j}, \ldots, \alpha_{M,j}$ must be 0, otherwise the objective function goes to infinity, which means the j-th candidate is not selected. The new formulation appears to be similar to (6), but we can prove that it is a convex problem and thus guarantees a global optimal solution.

THEOREM 1. *The problem (7) is convex w.r.t.* $\boldsymbol{\beta}, \{\boldsymbol{\alpha}_i\}$.

PROOF. In the objective function, the first term is a square loss and the third term is an ℓ_1-norm, both are known to be convex. Since a summation of convex functions is also convex, in the following we only need to prove that $S_j = \sum_i \alpha_{i,j}^2 / \beta_j$ is convex too. A sufficient and necessary condition of S_j being convex is that its Hessian ∇S_j is positive semidefinite, therefore we compute the second-order derivative terms

$$\frac{\partial S_j}{\partial \beta_j \partial \beta_j} = \frac{2}{\beta_j^3} \boldsymbol{\alpha}_{\cdot,j}^\top \boldsymbol{\alpha}_{\cdot,j}$$

$$\frac{\partial S_j}{\partial \alpha_{i,j} \partial \alpha_{i,j}} = \frac{2}{\beta_j} = \frac{2}{\beta_j^3} \boldsymbol{\psi}_{i,j}^\top \boldsymbol{\psi}_{i,j}$$

$$\frac{\partial S_j}{\partial \alpha_{i,j} \partial \beta_j} = -\frac{2\alpha_{i,j}}{\beta_j^2} = -\frac{2}{\beta_j^3} \boldsymbol{\alpha}_j^\top \boldsymbol{\psi}_{i,j}$$

where $\boldsymbol{\alpha}_{\cdot,j} = (\alpha_{1,j}, \ldots, \alpha_{M,j})^\top$, and $\boldsymbol{\psi}_{i,j}$ is a vector of length M, whose i-th element is $-\beta_j$ and the rest are zeros. Therefore we have

$$\nabla S_j = \frac{2}{\beta_j^3} \boldsymbol{\Psi}_j^\top \boldsymbol{\Psi}_j \tag{8}$$

where $\boldsymbol{\Psi}_j = (\boldsymbol{\alpha}_{\cdot,j}, \boldsymbol{\psi}_{1,j}, \ldots, \boldsymbol{\psi}_{M,j})$. Given the condition $\beta_j \geq 0$ and thus $\beta_j^3 \geq 0$, therefore $\nabla S_j \succeq 0$. The proof is completed. □

By noticing the inequality

$$\frac{\sum_{i=1}^{M} \alpha_{i,j}^2}{|\beta_j|} + \gamma |\beta_j| \geq 2 \sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2}, \tag{9}$$

whose equality holds only if $\beta_j^2 = \frac{1}{\gamma} \sum_{i=1}^{M} \alpha_{i,j}^2$, we obtain a necessary condition for the optimal solution of (7). We plug this condition into (7) and find that the problem (7) is equivalent to

$$\min_{\boldsymbol{\alpha}_i \in \mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_\mathcal{C}^\top \boldsymbol{\alpha}_i\|^2 + 2 \sum_{j=1}^{N} \sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2} \tag{10}$$

Though (7) and (10) are equivalent, we find it convenient to implement an iterative algorithm to find the local optimum of (7), which is also the global optimum, without accessing any optimization package. The algorithm is based on the observation that the update of $\boldsymbol{\beta}$ or $\boldsymbol{\alpha}_i$ while fixing the other has an analytical solution.

$$\boldsymbol{\alpha}_i = (\operatorname{diag}(\boldsymbol{\beta})^{-1} + \mathbf{X}_\mathcal{C} \mathbf{X}_\mathcal{C}^\top)^{-1} \mathbf{X}_\mathcal{C} \mathbf{x}_i, \quad i = 1, \ldots, M \tag{11}$$

$$\beta_j = \sqrt{\frac{1}{\gamma} \sum_{i=1}^{M} \alpha_{i,j}^2}, \quad j = 1, \ldots, N. \tag{12}$$

The iterations proceed as described in Algorithm 2. Similar to Algorithm 1, the method is extremely easy to implement, without requiring any optimization package. The major computational cost comes from the update of $\boldsymbol{\alpha}_i$, which can be greatly simplified by applying the Woodbury identity if the dimensionality of data is smaller than the size of candidates, which is the case in our experiments since we often reduce the dimensionality of data before the data selection. The computational cost can be further reduced by considering smaller sizes of $\mathbf{X}_\mathcal{C}$ and $\mathbf{X}_\mathcal{P}$ that are random subsets from all available unlabeled data.

3.4 The Control of Sparsity

Compared with the greedy procedure described in Algorithm 1, we lose the direct control on the budget $|\mathbf{X}_\mathcal{A}| = K$. A detailed relationship between γ and the resultant K needs to be investigated, e.g., computing the whole solution path,

Algorithm 2 Convex TED

```
Require: candidates $\mathbf{X}_{\mathcal{C}}$, unlabeled data $\mathbf{X}_{\mathcal{P}}$, $\gamma > 0$;
initialize $(\alpha_{i,j})$;
repeat
    $\beta_j \leftarrow \sqrt{\frac{1}{\gamma}\sum_{i=1}^{M}\alpha_{i,j}^2}$ for $j = 1, \ldots, N$;
    $\boldsymbol{\alpha}_i \leftarrow (\mathrm{diag}(\boldsymbol{\beta})^{-1} + \mathbf{X}_{\mathcal{C}}\mathbf{X}_{\mathcal{C}}^{\top})^{-1}\mathbf{X}_{\mathcal{C}}\mathbf{x}_i$, for $i = 1, \ldots, M$;
until converge;
$\mathbf{X}_{\mathcal{A}} \leftarrow \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_{\mathcal{C}}, \beta_j \neq 0\}$;
return $\mathbf{X}_{\mathcal{A}}$
```

which is however not the scope of the current paper. We conjecture that the sparsity is almost monotonic with respect to the regularization weight γ, so we can do line search to find an appropriate γ that selects roughly K examples. We provide Theorem 2 to show that the range of γ is upper bounded by a value γ_{max}, by checking the condition of producing *at least one non-zero element* in $\boldsymbol{\beta}$. This result can be used to narrow down the range of our line search.

THEOREM 2. *A necessary condition for the cardinality constraint* $|\boldsymbol{\beta}|_0 \geq 1$ *is*

$$\gamma \leq \gamma_{max} = \max_{j \in \mathcal{C}} \sum_{i \in \mathcal{P}} (\mathbf{x}_i^{\top}\mathbf{x}_j)^2 \tag{13}$$

PROOF. Suppose that γ is sufficiently large so that $\beta_j = 0$ and thus $\alpha_{i,j} = 0$ for all i and j. We compute the partial derivatives of the two costs in (10)

$$\mathbf{d}_1 = \frac{\partial \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\boldsymbol{\alpha}_i\|^2}{\partial \boldsymbol{\alpha}_{\cdot,j}} = -2 \cdot (\mathbf{x}_1^{\top}\mathbf{x}_j, \ldots, \mathbf{x}_M^{\top}\mathbf{x}_j)$$

$$\mathbf{d}_2 = \frac{\partial 2 \sum_{j=1}^{N} \sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2}}{\partial \boldsymbol{\alpha}_{\cdot,j}} = \frac{2\sqrt{\gamma}}{\sqrt{\sum_i \alpha_{i,j}^2}}(\alpha_{1,j}, \ldots, \alpha_{M,j})$$

where $\boldsymbol{\alpha}_{\cdot,j} = (\alpha_{1,j}, \ldots, \alpha_{M,j})$. It is easy to see that $\|\mathbf{d}_2\|^2$ equals to a constant $= 4\gamma$. The first derivative tends to pull $\boldsymbol{\alpha}_{\cdot,j}$ away from the origin while the second tends to push $\boldsymbol{\alpha}_{\cdot,j}$ toward the origin. Thus in order to pull out nonzero elements of $\alpha_{i,j}$ or β, there should be $\|\mathbf{d}_2\| \leq \|\mathbf{d}_1\|$, which completes the proof. □

However the line search is still too costly, because it has to solve an optimization problem (7) at each step of changing γ. In this paper we develop a practical heuristic to obtain result subject to the budget constraint $|\boldsymbol{\beta}|_0 = K$. The basic idea is that, instead of changing γ to enforce the desired sparsity of the result, we can change the problem to make it indeed sparse, and ensure that a result with the desired sparsity can be produced with a high probability under a fixed roughly chosen γ. If the intrinsic dimensionality of data in $\mathbf{X}_{\mathcal{P}}$ is K', we know that it requires at least K' linearly independent factors whose linear combination can sufficiently approximate every data example in $\mathbf{X}_{\mathcal{P}}$. Therefore K' offers a *lower-bound* for the sparsity $|\boldsymbol{\beta}|_0 = K$ of the result. If we increase the lower bound K', the obtained K is likely to be increased. This explains why empirically we found it very effective to control the sparsity K of result by changing the intrinsic dimensionality K' of data using principal component analysis (PCA). The detailed steps are the following

1. Perform singular value decomposition (SVD) on the unlabeled data

   $$\mathbf{X}_{\mathcal{P}} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\top} \tag{14}$$

   where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{K'})$.

2. Project data into the K' principal space

   $$\tilde{\mathbf{X}}_{\mathcal{P}} = \mathbf{X}_{\mathcal{P}}\mathbf{V} \tag{15}$$

   $$\tilde{\mathbf{X}}_{\mathcal{C}} = \mathbf{X}_{\mathcal{C}}\mathbf{V} \tag{16}$$

3. Replace $\mathbf{X}_{\mathcal{P}}$ and $\mathbf{X}_{\mathcal{C}}$ with $\tilde{\mathbf{X}}_{\mathcal{P}}$ and $\tilde{\mathbf{X}}_{\mathcal{C}}$ in problem (7), and solve it to obtain $\boldsymbol{\beta}^*$.

4. Retrieve the data examples from $\mathbf{X}_{\mathcal{C}}$, whose corresponding weights are ranked among the top K elements of $\boldsymbol{\beta}^*$, and form the active set $\mathbf{X}_{\mathcal{A}}$.

Note that we use the low-dimension data for data selection only, while use the original full-dimension data for training the classifier. In our experiments we found it effective to use a simple linear relationship

$$K = \rho K' \tag{17}$$

to control the sparsity, where ρ can be empirically estimated from data (e.g., see Section 4). We simply choose the regularization parameter γ among $\{0.001, 0.01, 0.1\} \times \gamma_{max}$. This approach is computationally much cheaper than line search, and is also cheaper than solving the original (7) due to the reduced dimensionality of data. We note that a theoretical linear relationship between the resultant sparsity K and the data dimensionality K' has been suggested by Donoho [4] as a condition of equivalence between ℓ_0-norm and ℓ_1-norm minimization. A detailed connection to this theoretical work is not the purpose of the present paper, we will instead provide some empirical justification in our experiments.

4. EXPERIMENTS

4.1 Data and Experimental Settings

In this section we evaluate the proposed convex TED algorithm for text categorization. Our empirical study was conducted based on two real-world text corpora. Our first data set is a subset of Newsgroup corpus, which contains 3970 documents with 8014 dimensional TFIDF features. This data set covers four categories: ‘autos’, ‘motorcycles’, ‘baseball’, ‘hockey’, each with 988, 993, 992 and 997 documents respectively. The other data set is a subset of the RCV1-v2 text data set, provided by Reuters and corrected by Lewis et al. [9]. The data set contains the information of topics, regions and industries for each document and a hierarchical structure for topics and industries. A set of 10000 documents is chosen for our experiments, including categories ‘C15’, ‘MCAT’, ‘GCAT’, and ‘CCAT’, each with 1826, 2477, 2999, and 4671 documents respectively. In this data set we use 9705 dimensional TFIDF features.

We conduct *one-against-all* classification for each category and thus treat the problem as binary classification, i.e., $y = \{-1, 1\}$, where documents from the target category are labeled as positive one, while those not belonging to this category are labeled as negative one. In the one-against-all setting, each binary classification task has unbalanced data, i.e., 18% $\sim$ 25% examples are positive and the rest negative. Because each classification task is unbalanced, classification accuracy rate may not be suitable to measure the performance. In the experiments we use the AUC score, i.e., *area under the Receiver Operating Characteristic (ROC) curve*, to measure the overall classification performance. Another reason is that ROC curves use true positive rates and false positive rates, which is closely related to precision and recall commonly used in IR tasks.

In each run of our experiments, an active learning method is applied to select a given number K of training examples, $K = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, then a classifier is trained on these examples with their labels. The trained classifier is then used to predict the class labels of the remaining examples, and an AUC score is computed based on the results. In order to randomize the experiments, in each run of experiments we restrict the training examples to be selected from a random candidate set of 50% of the total data. Therefore for each combination of active learning method and a number K, we compute the mean and standard error based on 10 randomized experiments. In this paper, we evaluate and compare five active learning methods,

- **Random Sampling** method uniformly selects examples as training data. We use this method as the baseline for active learning.

- **K-Means** method performs K-Means clustering algorithm on the data first, then selects the centroids of each cluster as the training data. The number of clusters is exact the number of training examples to be selected.

- **Simple Margin** method is a method in [16]. This method selects the example closest to the current decision boundary of the classifier, which is a usual SVM using the hinge loss.

- **Sequential TED** method, as described in Algorithm 1, which greedily selects the example which minimize the loss.

- **Convex TED** method, as described in Algorithm 2, is the new method proposed in this paper.

We note that all the methods use least-squares SVM (LSSVM) as the base classification method, except the Simple Margin method that uses hinge-loss SVM. In all the experiments we fix the parameters as $\mu = 0.01$ and $\gamma = 0.1\gamma_{max}$.

4.2 Sparsity of Results

We first empirically illustrate some properties of our algorithm in the aspect of achieving sparse results. Due to the space limitation, all the results shown here are based on the Newsgroup data. However similar phenomena can be observed on the other data set as well. Figure 1 shows that the achieved sparsity K can be controlled by either changing the regularization parameter γ, or changing the dimensionality K' of data as suggested in Section 3.4. However, as suggested in Figure 1-(a), the resultant sparsity K has a monotonic but nonlinear dependence on the parameter γ, which makes the line search difficult to ensure the sparsity budget. In contrast, as suggested by Figure 1-(b), we empirically found the simple linear connection between K and the data dimensionality K' (in this case $\rho = 1.61$). This provides a hint for us to efficiently achieve the desired sparsity.

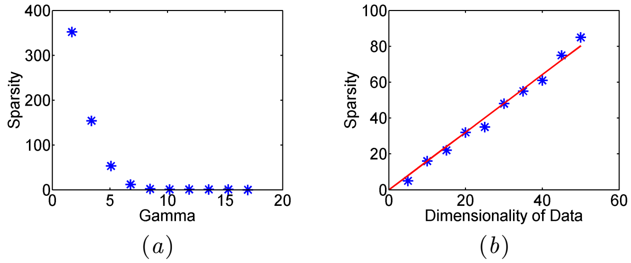


Figure 1: Control the sparsity of optimization results: (a) complex non-linear dependence on γ vs. (b) simple linear relationship with the data dimensionality K'.

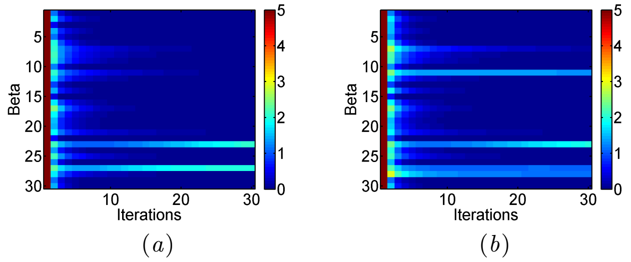


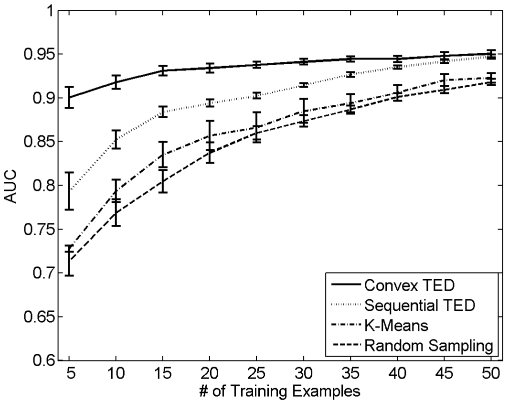
Figure 2: Updates of $\boldsymbol{\beta}$ (30 dimensions in this case) over iterations (a) $K' = 5$ (b) $K' = 10$.

In fact, since the examples are selected based on the ranking of elements in β^*, we only require the achieved sparsity to be approximately close to K. Therefore ρ does not have to be very accurate. In our experiments we simply fix $\rho = 1$.

In the next, we use an example to show that our formulation can indeed achieve sparse results across iterations of Algorithm 2. For illustration purpose, we restrict the candidate set $\mathbf{X}_\mathcal{C}$ to be 30 random examples from the 3970 documents, therefore $\boldsymbol{\beta}$ is a 30-dimension vector of non-negative numbers, which are initialized to be $\beta_j = 5,\ j = 1, \ldots, 30$. As shown in Figure 2, in both cases of $K' = 5$ and $K' = 10$, most of the elements in $\boldsymbol{\beta}$ vanish to zero quickly over iterations. Eventually only 2 and 5 elements survive to be nonzero at the 30-th iteration. In general, we found the algorithm locate the active set within hundreds of iterations given many thousands of candidates.

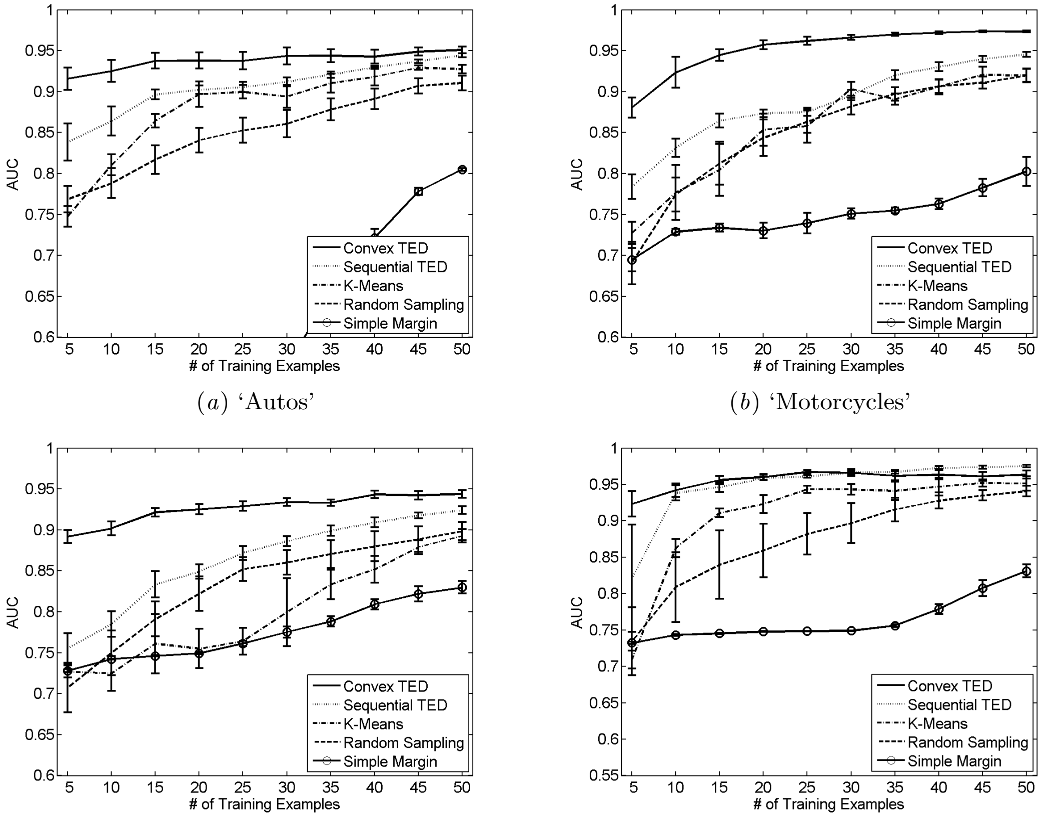
4.3 Performance Evaluation

Newsgroup: We compute the AUC scores for each experiments of a given method and a given K, and average them over all the categories and random trials to obtain the overall performance of each pair of method and K. The results are shown in Table 1 with a figure on the left and a table showing the numbers on the right. A little bit surprisingly, the K-Means approach only performs slightly better than the Random Sampling approach. Sequential TED exhibits performances better than K-Means and Random Sampling, but worse than the Convex TED method. The new algorithm produces very impressive results in this case: classifiers trained on only 5 training examples give the AUC score 90% in average, while Random Sampling needs almost 10 times of that training size to reach the same level of accuracy.



K	convex TED	sequential TED	K-Means	Random Samp.
5	0.900 ± 0.012	0.793 ± 0.021	0.728 ± 0.004	0.714 ± 0.017
10	0.918 ± 0.008	0.852 ± 0.010	0.794 ± 0.013	0.769 ± 0.015
15	0.931 ± 0.005	0.884 ± 0.006	0.835 ± 0.015	0.805 ± 0.013
20	0.934 ± 0.005	0.894 ± 0.004	0.857 ± 0.017	0.838 ± 0.012
25	0.938 ± 0.003	0.902 ± 0.003	0.866 ± 0.017	0.860 ± 0.008
30	0.941 ± 0.003	0.914 ± 0.002	0.885 ± 0.014	0.874 ± 0.006
35	0.944 ± 0.003	0.927 ± 0.003	0.894 ± 0.010	0.887 ± 0.005
40	0.944 ± 0.004	0.935 ± 0.002	0.906 ± 0.009	0.901 ± 0.004
45	0.948 ± 0.004	0.942 ± 0.002	0.920 ± 0.007	0.909 ± 0.004
50	0.950 ± 0.004	0.947 ± 0.002	0.923 ± 0.005	0.918 ± 0.003

Table 1: Overall performance on Newsgroup data. Both of the figure and the table show the mean and the standard error of AUC score over 10 random trials.



(a) 'Autos'

(b) 'Motorcycles'

(c) 'Baseball'

(d) 'Hockey'

Figure 3: Classification performance on different categories of Newsgroup data set. For category 'Autos', the curve of Simple Margin cannot be completely plotted due to its very poor performance in this case.

Note that here we compare only the methods whose data selection is *label-independent*. The Simple Margin method selects the next data based on the current classifier, which is trained on the previously seen labels of examples. Therefore, for each category the method selects a *different* set of training data that are largely non-overlapped, while those data-independent methods select a *common* set of data for training classifiers of all the categories. In order to put the performance of each binary classifier in a comparable ground, each learner is supposed to pick up the same number K of examples, which however makes the comparison a bit unfair for the multi-category case, because then Simple Margin actually employs $4K$ (if no overlap) examples in total while other methods use only K examples.

Nevertheless we can still make a comparison in the binary classification case for each individual category, whose results are plotted in Figure 3. For the 3 categories 'Autos', 'Motorcycles' and 'Baseball', Convex TED method outperforms the second best — Sequential TEC — by a large margin, and performs similarly on the category 'Hockey'. The Simple Margin method works surprisingly bad in this data set. We conjecture that the method is trapped to find outlier or untypical examples that are likely to stay close to the boundary of classifiers. In contrast, the TED methods set the optimization goal as finding data to well preserve the rest of other data, and are thus unlikely to find those outliers.

RCV: The overall performances of different methods except Simple Margin on the RCV data are presented in Table 2. In this case the K-Means method performs very close to the Sequential TED method, except for the case of $K = 5$ where the AUC of Sequential TED is 75.3% while that of K-Means is 65.9%. The Convex TED method demonstrates the best performance in all the cases when less than 40 examples are selected. Especially the advantage is notable when $K = 5, 10, 15$ and 20. Furthermore, as shown in Figure 4 all the methods including Simple Margin are compared on the bases of individual categories. We find that Simple Margin performs quite good on the categories of 'C15' and 'CCAT', while very poor on the category 'GCAT'. Sequential TED and K-Means behave very closely. Random Sampling shows fair performance except on 'C15'. Convex TED outperforms Sequential TED and exhibits good results in all the cases, especially on the category 'GCAT' where Simple Margin clearly fails.

Label-independent vs. Label-dependent: The experiment results, especially regarding those of Convex TED and Simple Margin methods, raise an interesting question: what are the pros and cons of label-dependent/independent active learning algorithms? We feel that label-independent active learning, e.g., TED, tends to be somewhat more generative than discriminative, because it explores the distribution of unlabeled data and is less prone to untypical patterns or outliers. Moreover, as we discussed already, label-independent approaches seem to be cheaper for *active learning with multiple categories*, a topic rarely touched by the previous research. In practice, label-independent active learning should be useful in the early stage when very little labels are known. Our experiments showed that the advantage of Convex TED is particularly prominent when the training size is very small, the task involves multiple classes, or outliers are present. On the other hand, label-dependent active learning is naturally compelling, because it uses the information of labels. For example, the Simple Margin method performed very well for the category 'C15', which turns out to be the smallest category among the four. It could be understood that a label-dependent approach is good at finding small categories, as the information of known labels helps to push the classification boundary to shrink and be more focusing. However, these methods are usually too greedy, and sometimes prone to untypical data, as what we observed on the Newsgroup data. This analysis suggests a future direction of combining two kinds of approaches to make active learning less greedy and meanwhile discriminative.

5. CONCLUSION

In this paper we developed a non-greedy active learning algorithm by extending the framework of transductive experimental design (TED). Unlike the previous sequential greedy algorithm, the new formulation can simultaneously select multiple data examples at a time with a global optimum. We proposed an iterative algorithm that does not require to apply any non-trivial optimization technique. Our initial experiments demonstrated that the non-greedy solution outperforms the greedy approach, and produced good performances for active text categorization, especially at the initial phase. In the future, it is interesting to combine the strengths of convex TED and label-dependent methods to develop active learning approaches that are label-dependent and also less greedy.

6. REFERENCES

[1] A. C. Atkinson and A. N. Donev. *Optimum experiment designs*. Oxford Statistical Science Series. Oxford University Press, 1992.

[2] O. Chapelle. Active learning for Parzen window classifier. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 49–56, 2005.

[3] D. Cohn and Z. Ghahramani. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[4] D. Donoho. For most large underdetermined systems of linear equations, the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6), 2006.

[5] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[6] C. Guestrin, A. Krause, and A. Singh. Near-optimal sensor placements in gaussian processes. In *Proc. of the International Conference on Machine Learning (ICML)*, 2005.

[7] X. He, W. Min, D. Cai, and K. Zhou. Laplacian optimal design for image retrieval. In *ACM SIGIR Conference*, 2007.

[8] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *International Conference on Machine Learning (ICML)*, 2006.

[9] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 2005.

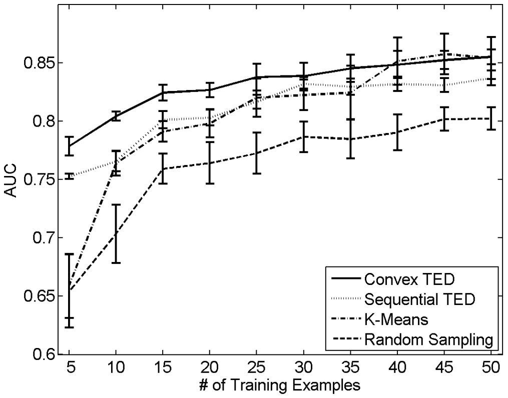
[10] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[11] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[12] A. Schein and L. Ungar. Optimality for active learning of logistic regression classifiers. Technical Report Technical Report MS-CIS-04-07, The University of Pennsylvania, Department of Computer and Information Science, 2004.

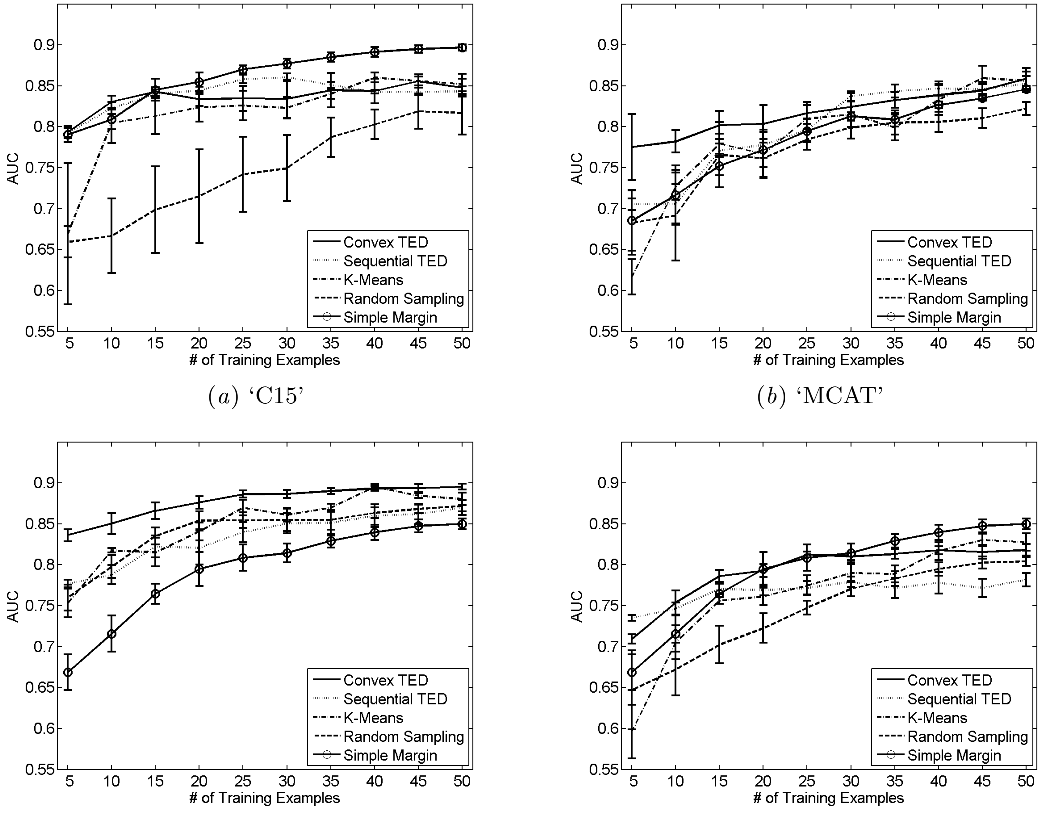
[13] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, 2000.

[14] J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 1999.



K	convex TED	sequential TED	K-Means	Random Samp.
5	0.778 ± 0.008	0.753 ± 0.002	0.659 ± 0.028	0.654 ± 0.031
10	0.804 ± 0.004	0.766 ± 0.009	0.764 ± 0.010	0.704 ± 0.025
15	0.824 ± 0.007	0.801 ± 0.007	0.791 ± 0.009	0.759 ± 0.013
20	0.827 ± 0.006	0.803 ± 0.007	0.798 ± 0.012	0.764 ± 0.018
25	0.838 ± 0.011	0.817 ± 0.006	0.820 ± 0.016	0.772 ± 0.018
30	0.839 ± 0.011	0.832 ± 0.006	0.822 ± 0.013	0.786 ± 0.013
35	0.845 ± 0.012	0.829 ± 0.007	0.824 ± 0.023	0.785 ± 0.017
40	0.848 ± 0.012	0.832 ± 0.006	0.851 ± 0.020	0.790 ± 0.015
45	0.852 ± 0.008	0.831 ± 0.006	0.858 ± 0.017	0.802 ± 0.010
50	0.855 ± 0.006	0.837 ± 0.006	0.854 ± 0.018	0.802 ± 0.010

Table 2: Overall performance on RCV data. Both of the figure and the table show the mean and the standard error of AUC score over 10 random trials.



(a) 'C15'

(b) 'MCAT'

(c) 'GCAT'

(d) 'CCAT'

Figure 4: Classification performance on different categories of RCV data set.

[15] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B*, 58(1), 1996.

[16] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 2001.

[17] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *International Conference on Machine Learning (ICML)*, 2006.

[18] J. Zhang and Y. Yang. Robustness of regularized linear classifcation methods in text categorization. In *The 26th Annual International SIGIR Conference (SIGIR'99)*, 2003.

[19] T. Zhang and F. J. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, (4):5–31, 2001.

[20] W. V. Zhang, X. He, B. Rey, and R. Jones. Query rewritting using active learning for sponsored search. In *ACM SIGIR Conference*, 2007.