SED: Supervised Experimental Design and Its Application to Text Classification

Yi Zhen and Dit-Yan Yeung
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong, China
{yzhen, dyyeung}@cse.ust.hk

ABSTRACT

In recent years, active learning methods based on experimental design achieve state-of-the-art performance in text classification applications. Although these methods can exploit the distribution of unlabeled data and support batch selection, they cannot make use of labeled data which often carry useful information for active learning. In this paper, we propose a novel active learning method for text classification, called supervised experimental design (SED), which seamlessly incorporates label information into experimental design. Experimental results show that SED outperforms its counterparts which either discard the label information even when it is available or fail to exploit the distribution of unlabeled data.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Experimental Design*; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms

Algorithms, Theory

Keywords

Active Learning, Supervised Experimental Design, Text Classification, Convex Optimization

1. INTRODUCTION

There has been a long tradition of research on text classification in both the information retrieval and machine learning communities. In order to learn a good text classifier, a large number of labeled documents are often needed for classifier training. However, labeling documents always needs domain knowledge and thus is difficult, time consuming and costly. On the other hand, it is much easier to obtain a large number of unlabeled documents, such as web pages, newspapers and journal articles. In recent years, a new approach called *active learning* [1, 3, 5, 6, 9, 11, 13, 14, 15, 16, 18, 20, 25] has been developed in the machine learning community with the goal of reducing the labeling cost by identifying and presenting the most informative examples from the unlabeled examples for the human experts to label.

Although a lot of work has been done in active learning research, most of the existing active learning methods are still far from satisfactory with apparent shortcomings. In particular, many methods only take into consideration partial information to determine the informativeness of examples. Some methods consider information conveyed by the class boundaries, some consider information conveyed by the distribution of unlabeled data, and some consider the disagreement between learners when multiple learners are involved. Unfortunately, none of these methods is consistently better than others in all situations. Another drawback is that most active learning algorithms select only one example at a time for labeling. Compared with a batch approach [7, 9, 10] which selects multiple examples in one iteration, this greedy incremental approach is at best suboptimal and is not suitable for large-scale and parallel computing applications.

Experimental design [2, 20, 21], which is one of the state-of-the-art active learning approaches for text classification, can effectively exploit the distribution of unlabeled data while supporting batch selection at the same time. Despite their appealing properties, existing methods based on experimental design cannot make use of label information even when labeled data are available. Thus, these methods are intrinsically unsupervised in nature.

In this paper, we propose a novel batch mode active learning algorithm, called supervised experimental design (SED), which incorporates label information into the experimental design procedure. SED is a supervised extension of experimental design with a new regularization term that incorporates label information added to the objective function. To the best of our knowledge, no work has been done so far to utilize label information in the experimental design procedure. Some favorable properties of SED are highlighted here:

- To the best of our knowledge, SED is the first work that incorporates label information into experimental design.
- SED outperforms (unsupervised) experimental design, which discards the label information even when it is available. This shows that label information does pro-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

vide useful information for active document selection and SED can utilize the information very effectively.

- SED outperforms margin-based active learning under highly unbalanced data distributions which are often encountered in practice.

- SED is convex and thus global optimality can be guaranteed.

The remainder of this paper is organized as follows. In Section 2, we will introduce the notations and some related work. In Section 3, we first introduce transductive experimental design and then present our SED model and algorithm in detail. Extensive empirical studies conducted on two real-world text corpora are presented in Section 4. Section 5 concludes our paper.

2. NOTATIONS AND RELATED WORK

Throughout this paper, we use boldface uppercase letters (e.g. $\mathbf{X}$) to denote matrices and boldface lowercase letters (e.g. $\mathbf{x}$) to denote vectors. We use $\mathrm{tr}(\mathbf{X})$ to denote the trace of $\mathbf{X}$ and $\mathbf{X}^T$ to denote its transpose. Moreover, we use calligraphic letters (e.g. $\mathcal{A}$) to denote sets and $|\mathcal{A}|$ to denote the size of $\mathcal{A}$.

Given the whole data set represented as $\mathbf{X}_{\mathcal{P}} \in \mathbb{R}^{M\times D}$ or $\mathcal{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, in which each data point $\mathbf{x}_i$ is a $D \times 1$ vector, a generic active learning problem [4, 11] can be defined as selecting a subset of unlabeled data points from a *candidate set* $\mathbf{X}_{\mathcal{C}} \in \mathbb{R}^{N\times D}$ or $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, such that if the selected data points are labeled and added to the training set for re-training the classifier, the improvement of the classifier will be maximized. We call the subset of selected data the *active set* and denote it as $\mathbf{X}_{\mathcal{A}} \in \mathbb{R}^{K\times D}$ or $\mathcal{A} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$.[1]

The promise of active learning is appealing because it can help to alleviate the labeled data deficiency problem commonly encountered in many supervised learning applications. Existing active learning algorithms for text classification either select the most uncertain data given the current classifier [11], select the data with the smallest margin [18], select the data on which multiple classifiers disagree most with each other [5, 14, 17], or select the data that optimize some information gain [6, 13, 16, 25].

Closely related to active learning is experimental design in statistics [2]. Conventionally, experimental design considers the problem of learning a predictive function $f(\mathbf{x})$ from experiment-measurement pairs $(\mathbf{x}_i, y_i)$. Given that conducting an experiment is expensive, experimental design seeks to select the most informative experiments to conduct such that the number of measurements needed can be reduced.

Traditional experimental design considers the following linear regression model:

$$y = \mathbf{w}^T\mathbf{x} + \epsilon,$$

where y is the measurement, $\mathbf{x}$ is the $D \times 1$ feature vector of the experiment, $\mathbf{w}$ is the $D \times 1$ model parameter vector and ϵ is the noise term.

[1]The reader should note that points in different sets with the same index are not necessarily the same point, although we require that the points in $\mathcal{C}$ should appear in $\mathcal{P}$ and the points in $\mathcal{A}$ should appear in $\mathcal{C}$.

Given a set of labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^M$, the maximum likelihood estimate (MLE) of the model parameter vector $\mathbf{w}$ can be obtained by minimizing the residual sum of squares:

$$\begin{aligned}\hat{\mathbf{w}} &= \operatorname*{argmin}_{\mathbf{w}} \Big\{ J(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{M} (\mathbf{w}^T\mathbf{x}_i - y_i)^2 \Big\} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \end{aligned} \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]^T$ is a matrix of the labeled data and $\mathbf{y}$ is a vector of the corresponding target outputs.[2]

If we put a spherical Gaussian prior on the noise ϵ, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$, it can be proved easily that $\hat{\mathbf{w}}$ is an unbiased estimate of $\mathbf{w}$ with covariance:

$$\begin{aligned}\mathrm{cov}[\hat{\mathbf{w}}] &= \mathrm{cov}\big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\big] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{cov}[\mathbf{y}]\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

Traditional experimental design aims at minimizing the covariance of $\hat{\mathbf{w}}$, which characterizes the model uncertainty in some sense. Three criteria have been commonly used in the literature:

- D-optimal design: minimizing the determinant of $\mathrm{cov}[\hat{\mathbf{w}}]$;
- A-optimal design: minimizing the trace of $\mathrm{cov}[\hat{\mathbf{w}}]$;
- E-optimal design: minimizing the maximum eigenvalue of $\mathrm{cov}[\hat{\mathbf{w}}]$.

Recently, Yu *et al.* [20] proposed a method, called *transductive experimental design* (TED), which selects the most informative examples by reducing the model uncertainty on all of the unlabeled data and thus effectively exploits the distribution of the unlabeled data. He *et al.* [8] applied similar ideas to content-based image retrieval (CBIR), where a Laplacian regularization term is added and then the model uncertainty, represented by a new covariance matrix, considers the smoothness among data points.

Despite the appealing properties which include clear mathematical formulation and the ability of batch selection, algorithms based on experimental design often have to deal with combinatorial complexity and are NP-hard. Since the optimization problems involved are non-convex, the solutions obtained may correspond to poor local minima. To address this problem, some approximation methods based on convex relaxation have been developed [21, 23].

3. SUPERVISED EXPERIMENTAL DESIGN

Existing active learning methods based on experimental design, such as TED, are formulated under the setting that all available data are unlabeled. As such, they cannot make use of the label information even when it is available.

Since label information has been found very useful to example (or document) selection [5, 11, 14, 16, 18], incorporating label information into the example selection procedure of experimental design is a very worthwhile direction to explore.

In this section, we first briefly review TED in Section 3.1 and then present our method, SED, in Section 3.2. The algorithm for SED will be summarized in Section 3.3 and its complexity analysis will be presented in Section 3.4.

[2]In the sequel, we will also refer to them as labels even though the term 'label' is more appropriately used for classification problems.

3.1 Transductive Experimental Design

TED [20] seeks to choose $\mathbf{X}_\mathcal{A}$ from $\mathbf{X}_\mathcal{C}$ such that a function f learned from $\mathbf{X}_\mathcal{A}$ has the smallest predictive variance on $\mathbf{X}_\mathcal{P}$. The goal can be achieved by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{X}_\mathcal{A}} \quad & \mathrm{tr}\left[\mathbf{X}_\mathcal{P}\left(\mathbf{X}_\mathcal{A}^T\mathbf{X}_\mathcal{A}+\mu\mathbf{I}\right)^{-1}\mathbf{X}_\mathcal{P}^T\right] \\ \text{s.t.} \quad & \mathbf{X}_\mathcal{A}\subset\mathbf{X}_\mathcal{C}, |\mathcal{A}|=K, \end{aligned} \tag{2}$$

where $\mathbf{I}$ is an identity matrix whose dimensionality is determined by the problem and K is the number of examples selected. The objective function may also be considered as model uncertainty over $\mathbf{X}_\mathcal{P}$. We note that it only depends on the input features of the training examples and thus is independent of the labels. This is because in the error function $J(\mathbf{w};\mathbf{X},\mathbf{y})$ of the linear regression model in Equation (1), the model parameter vector $\mathbf{w}$ is only coupled with the labels y_i linearly and hence a second derivative with respect to $\mathbf{w}$ makes all the y_i terms disappear.

Since the TED optimization problem is non-convex and can easily get stuck in local optima, Yu *et al.* [21] proposed a convex relaxation of TED (Convex TED). The optimization problem of Convex TED is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i} \quad & \textstyle\sum_{i=1}^M\left(||\mathbf{x}_i-\mathbf{X}_\mathcal{C}^T\boldsymbol{\alpha}_i||^2+\sum_{j=1}^N\frac{\alpha_{ij}^2}{\beta_j}\right)+\gamma||\boldsymbol{\beta}||_1 \\ \text{s.t.} \quad & \mathbf{x}_i\in\mathbf{X}_\mathcal{P}, \boldsymbol{\alpha}_i\in\mathbb{R}^N, \boldsymbol{\beta}\in\mathbb{R}^{N\times 1}, \boldsymbol{\beta}\geq\mathbf{0}, \end{aligned} \tag{3}$$

where the variables $\beta_j, j=1,\ldots,N$, control the inclusion of examples in $\mathbf{X}_\mathcal{C}$ into the training set $\mathbf{X}_\mathcal{A}$, the ℓ_1-norm $||\boldsymbol{\beta}||_1$ enforces the sparsity of $\boldsymbol{\beta}$, and α_{ij} denotes the jth element of $\boldsymbol{\alpha}_i$. According to [20, 21], TED and Convex TED tend to select examples representative of all the unlabeled data and hence exploit the distribution of the whole data space.

Since experimental design based methods do not use label information, we call them *unsupervised* active learning methods here. In the next subsection, we will present our supervised extension, SED, which can effectively utilize the available label information to select the most informative examples.

3.2 Supervised Experimental Design

Given a set of labeled data points (training set), we can learn a classifier f from the data. In a typical active learning setting in which labeled data are scarce, f may not be accurate enough and hence it is desirable to select some unlabeled data points for labeling to enlarge the training set. However, although f is not accurate enough, it still carries some useful information about the data points. Let $\mathbf{f}$ be a vector of decision values on the candidate set $\mathbf{X}_\mathcal{C}$ and $\tilde{\mathbf{f}}$ be the vector after taking the absolute value of each element of $\mathbf{f}$. For example, in support vector machine (SVM), $\tilde{\mathbf{f}}$ indicates the uncertainty of the current classifier about the labels of the examples. The smaller the jth element $\tilde{f}_j$ of $\tilde{\mathbf{f}}$ is, the less certain is the classifier about the example. Intuitively speaking, the most informative examples should be those with the smallest $\tilde{f}_j$ values.

Based on the above intuition and the formulation of TED, we can choose the most informative $\mathbf{X}_\mathcal{A}$ by solving the following optimization problem,

$$\begin{aligned} \min_{\mathbf{X}_\mathcal{A}} \quad & \mathrm{tr}\left[\mathbf{X}_\mathcal{P}\left(\mathbf{X}_\mathcal{A}^T\mathbf{X}_\mathcal{A}+\mu\mathbf{I}\right)^{-1}\mathbf{X}_\mathcal{P}^T\right]+\gamma\tilde{\mathbf{f}}_\mathcal{A} \\ \text{s.t.} \quad & \mathbf{X}_\mathcal{A}\subset\mathbf{X}_\mathcal{C}, |\mathcal{A}|=K, \end{aligned} \tag{4}$$

where $\tilde{\mathbf{f}}_\mathcal{A}$ is similar to $\tilde{\mathbf{f}}$ but is defined only on the active set $\mathbf{X}_\mathcal{A}$ and γ is a user-defined parameter controlling the contribution of model uncertainty due to the current labeled data. In other words, γ controls the contribution of label information.

Since the optimization problem in Equation (4) is NP-hard, non-convex and can easily get trapped in local minima, we borrow ideas from [21] to reformulate it in a convex form and define our SED problem as follows.

Definition 1. *Supervised Experimental Design (SED)*

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i} \quad & \textstyle\sum_{i=1}^M\left(||\mathbf{x}_i-\mathbf{X}_\mathcal{C}^T\boldsymbol{\alpha}_i||^2+\sum_{j=1}^N\frac{\alpha_{ij}^2}{\beta_j}\right)+\gamma_1||\boldsymbol{\beta}||_1+\gamma_2\boldsymbol{\beta}^T\tilde{\mathbf{f}} \\ \text{s.t.} \quad & \mathbf{x}_i\in\mathbf{X}_\mathcal{P}, \boldsymbol{\alpha}_i\in\mathbb{R}^N, \boldsymbol{\beta}\in\mathbb{R}^{N\times 1}, \boldsymbol{\beta}\geq\mathbf{0}. \end{aligned} \tag{5}$$

We can further prove that SED is a convex problem.

Theorem 1. *SED is convex w.r.t. $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}$.*

Proof. Let the objective function of SED be $F=F_1+F_2$, where $F_1=\sum_{i=1}^M\left(||\mathbf{x}_i-\mathbf{X}_\mathcal{C}^T\boldsymbol{\alpha}_i||^2+\sum_{j=1}^N\frac{\alpha_{ij}^2}{\beta_j}\right)+\gamma_1||\boldsymbol{\beta}||_1$ and $F_2=\gamma_2\boldsymbol{\beta}^T\tilde{\mathbf{f}}$. Because $\tilde{\mathbf{f}}$ is constant, F_2 is linear in $\boldsymbol{\beta}$. Thus F_2 is convex with respect to $\boldsymbol{\beta}$. Since F_1 is also convex with respect to $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}$[3] and F_1+F_2 is a convex combination of two convex functions F_1 and F_2, F is thus convex with respect to $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}$. This completes the proof. □

3.3 Algorithm

It is convenient to find the local optimum of Problem (5), which is also the global optimum, by updating $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}$ iteratively. More specifically, we can find the analytical solution for updating one variable while fixing the other as follows:

$$\beta_j=\sqrt{\frac{1}{\gamma_1+\gamma_2\tilde{f}_j}\sum_{i=1}^M\alpha_{ij}^2},\ j=1,\ldots,N, \tag{6}$$

$$\boldsymbol{\alpha}_i=(\mathrm{diag}(\boldsymbol{\beta})^{-1}+\mathbf{X}_\mathcal{C}\mathbf{X}_\mathcal{C}^T)^{-1}\mathbf{X}_\mathcal{C}\mathbf{x}_i,\ i=1,\ldots,M. \tag{7}$$

The proposed algorithm is summarized in Algorithm 1.

3.4 Complexity Analysis

The main computation of SED is to update $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}$. The time complexity of updating $\boldsymbol{\beta}$ (Equation (6)) is $O(MN)$ and that of updating $\{\boldsymbol{\alpha}_i\}$ (Equation (7)) is $O(N^3)$. Hence, the time complexity of one iteration is $O(N^3+MN)$. Though our algorithm converges very quickly in practice, it is interesting and worthwhile to explore more efficient techniques to solve the problem, and we leave it as future work.

4. EMPIRICAL ANALYSIS

We conduct several experiments to compare SED with some other related methods. We have the following questions in mind while designing and conducting the experiments:

1. How does SED perform when compared with other state-of-the-art active learning methods?
2. How effective is label information for experimental design?

[3]The proof can be found in [21].

Algorithm 1 Algorithm for SED

INPUT:
$\mathcal{L}^0$ – set of labeled data points
$\mathcal{U}^0$ – set of unlabeled data points
T – number of active learning iterations
K – number of examples selected in each iteration
for $t = 1$ to T **do**
  Train classifier f based on $\mathcal{L}^{t-1}$.
  Compute absolute decision values $\tilde{\mathbf{f}}$.
  Initialize $\{\boldsymbol{\alpha}_i\}$.
  repeat
    Fix $\{\boldsymbol{\alpha}_i\}$, update $\boldsymbol{\beta}$ using Equation (6).
    Fix $\boldsymbol{\beta}$, update $\{\boldsymbol{\alpha}_i\}$ using Equation (7).
  until converge w.r.t. objective value of Problem (5).
  Choose K examples with the largest $\boldsymbol{\beta}$ values into $\mathbf{X}_{\mathcal{A}}^{t-1}$ and request their labels.
  Update $\mathcal{L}^t \leftarrow \mathcal{L}^{t-1} \cup \mathbf{X}_{\mathcal{A}}^{t-1}$ and $\mathcal{U}^t \leftarrow \mathcal{U}^{t-1} \setminus \mathbf{X}_{\mathcal{A}}^{t-1}$.
end for
Train classifier f based on $\mathcal{L}^T$.
return f

3. How does varying the size of the candidate set affect the performance of SED?

These questions are answered in separate subsections: question 1 in Section 4.3, question 2 in Section 4.4.1, and question 3 in Section 4.4.2.

4.1 Data Sets

We conduct experiments on two public benchmark data sets. The first one is a subset of the Newsgroups corpus [21], which consists of 3,970 documents with TFIDF features of 8,014 dimensions. Each document belongs to exactly one of the 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 [12]. We randomly choose from the original data set 5,000 documents with TFIDF features of 6,451 dimensions. Each document belongs to at least one of the four categories: *CCAT*, *ECAT*, *GCAT* and *MCAT*. Some characteristics of the two data sets are summarized in Table 1 respectively.

Table 1: Characteristics of Data Sets

Data Sets	Category	# of Documents	# of Features
Newsgroups	Autos	988	8,014
	Motorcycles	993	
	Baseball	992	
	Hockey	997	
Reuters	CCAT	907	6,451
	ECAT	1,259	
	GCAT	1,524	
	MCAT	2,337	

4.2 Experimental Settings and Metrics

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 each Newsgroups data set and about 30% of the documents in each Reuters data set are positive). Note that a larger value of AUC indicates a better performance.

At each iteration of our experiments, an active learning method selects a set of $K = 5$ unlabeled examples from the candidate set. The selected examples are then labeled and added to the training set $\mathcal{L}$. The classifier is then trained on the expanded training set and used to predict the class labels of all documents. An AUC score is then computed based on the predictions. In order to randomize the experiments as well as to reduce the computational cost, we restrict the candidate set to randomly cover only a fraction of all the unlabeled documents. Ten different candidate sets are generated for each experiment and the average AUC value, together with the standard deviation, is reported.

We compare SED with four popular active learning methods for text classification:

- **Convex TED** [21], which is a convex relaxation of TED.
- **Sequential TED** [20], which sequentially selects examples using TED.
- **Margin**, which chooses the examples closest to the class boundary. This method implements the basic idea of [18] but uses the squared loss instead of the hinge loss. We use this method because it performs much better than [18] in practice.
- **Random Sampling**, which randomly selects examples from the candidate set.

We note that all the methods use kernel ridge regression, which is essentially equivalent to least squares SVM (LS-SVM), as the base classifier. LS-SVM has been reported to give state-of-the-art performance in text classification tasks [22, 24]. Since no labeled data exists in the beginning of each experiment, we use Convex TED to select the first $K = 5$ examples for SED and Margin.

4.3 Performance Evaluation

4.3.1 Comparison of Methods on Newsgroups Data

We first compare the five methods on the Newsgroups data set. For each method, we restrict the candidate set to cover 50% of the unlabeled data and set the parameters as $\mu = 0.01, \gamma_1 = 0.1\gamma_{\max}, \gamma_2 = 1$.[4]

The AUC values averaged over four binary classification tasks are reported in Table 2, where each row corresponds to one iteration. We use boldface numbers to indicate the best results among the five methods. It is obvious that SED consistently outperforms the other methods. To evaluate how significant SED outperforms other methods, we have conducted paired t-tests [19] on the results of SED and the second best method, Convex TED. The *p-value* is 2.37×10^{-5}, indicating that SED achieves a significantly better result. It is not surprising that Random Sampling performs the worst because the randomly selected examples may not provide much useful information to the classifier. We also note

[4] $\gamma_1 \leq \gamma_{\max} = \max_{j \in \mathcal{C}} \sum_{i \in \mathcal{P}} (\mathbf{x}_i^T \mathbf{x}_j)^2$ is a necessary condition for the cardinality constraint $\|\boldsymbol{\beta}\|_0 \geq 1$. The reader is referred to [21] for details.

Table 2: Comparison of Methods (in Average AUC) on Newsgroups Data

$\|\mathcal{L}\|$	SED	Convex TED	Sequential TED	Margin	Random Sampling
5	**0.8854**±0.0256	**0.8854**±0.0256	0.8195±0.0299	**0.8854**±0.0256	0.7138±0.0295
10	**0.9179**±0.0133	0.9057±0.0134	0.8501±0.0244	0.8800±0.0274	0.7703±0.0354
15	**0.9327**±0.0115	0.9186±0.0120	0.9023±0.0120	0.8914±0.0270	0.8006±0.0234
20	**0.9456**±0.0067	0.9244±0.0103	0.9219±0.0122	0.9027±0.0163	0.8261±0.0231
25	**0.9512**±0.0061	0.9361±0.0076	0.9304±0.0105	0.9115±0.0103	0.8460±0.0169
30	**0.9546**±0.0042	0.9407±0.0058	0.9362±0.0095	0.9171±0.0093	0.8639±0.0159
35	**0.9573**±0.0055	0.9446±0.0049	0.9406±0.0081	0.9212±0.0090	0.8782±0.0164
40	**0.9609**±0.0041	0.9471±0.0054	0.9434±0.0073	0.9262±0.0098	0.8904±0.0138
45	**0.9631**±0.0051	0.9493±0.0051	0.9460±0.0062	0.9321±0.0081	0.9009±0.0132
50	**0.9655**±0.0043	0.9514±0.0046	0.9486±0.0068	0.9363±0.0074	0.9076±0.0126

that the methods based on experimental design, i.e., Convex TED and Sequential TED, perform better than Margin. This indicates that exploiting the distribution of unlabeled data can provide more useful information than selecting only examples near the class boundary. We also observe that in the beginning of the learning procedure, examples selected by Margin actually degrade the performance. This is because the labeled data are scarce at that time and hence the class boundary learned by training on the labeled data is not accurate enough and hence may be misleading for document selection.

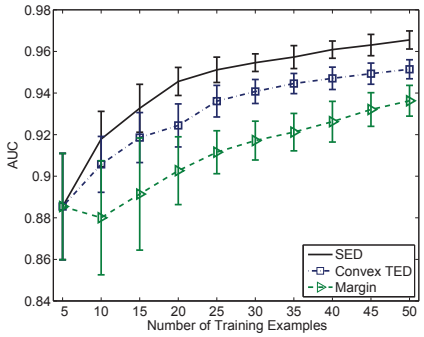


Figure 1: Learning Curves on Newsgroups Data

We plot the learning curves of SED, Convex TED and Margin in Figure 1. As we can see, SED performs better than its two counterparts by a large margin. This observation validates that considering label information and the distribution of unlabeled data together can provide more useful information for active selection than only considering either of them.

To further understand the properties of SED, we plot the learning curves of SED, Convex TED and Margin for four binary classification tasks in Figure 2. For three categories, i.e., *autos*, *baseball* and *hockey*, SED consistently outperforms the second best, Convex TED, by a large margin. For the *motorcycles* category, SED and Convex TED perform similarly. We note that Margin is consistently the worst with the largest variance for all tasks. We conjecture that Margin always selects the outliers, which stay close to the class boundary but are not useful to the learner. On the other hand, SED and TED can exploit the distribution of unlabeled data and hence have a smaller chance to select the outliers.

4.3.2 *Comparison of Methods on Reuters Data*

We now compare the five methods on the Reuters data set. Each candidate set covers 20% of the unlabeled documents. The parameters are set as $\mu = 0.01, \gamma_1 = 0.1\gamma_{\max}, \gamma_2 = 10$.

The AUC values averaged over the four tasks are reported in Table 3. Again the best results are shown in bold. As in the Newsgroups data set, SED significantly outperforms Convex TED (the *p-value* of paired t-test is 2.26×10^{-5}), validating the effectiveness of label information. It is interesting to find that Margin performs better than Convex TED. This can be attributed to two reasons. First, the data are very balanced in this data set and Margin selects the most discriminative examples without querying the outliers. Second, the representative examples selected by TED might not be as helpful as those discriminative ones. However, SED can take advantage of both criteria and always performs the best, especially in the early stage.

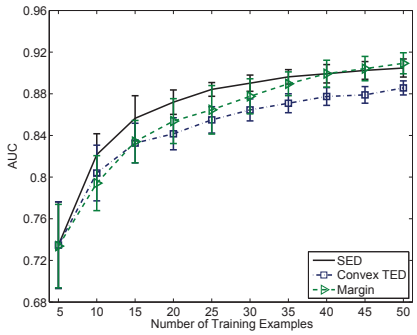
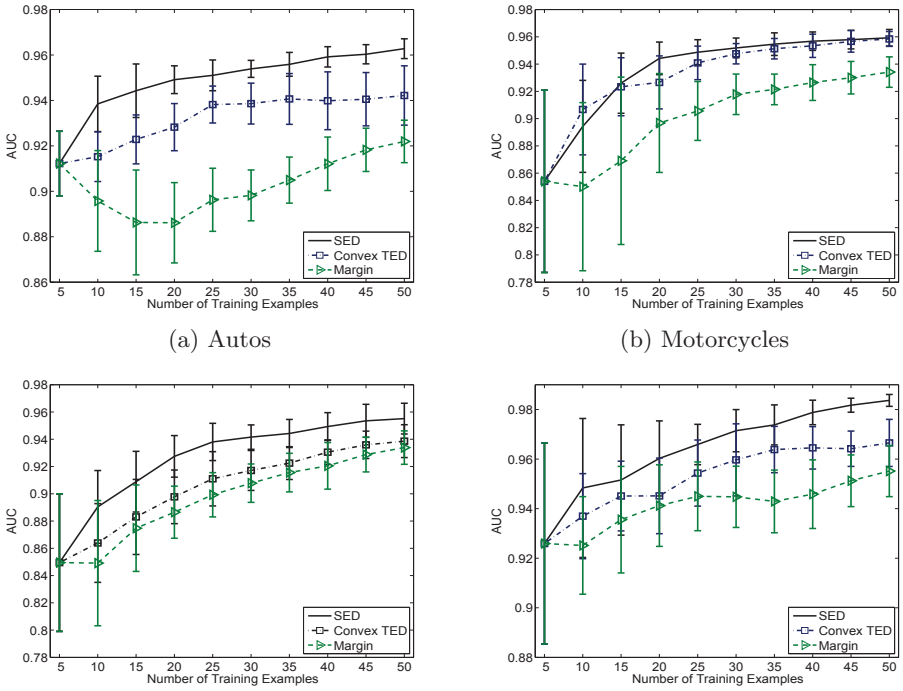


Figure 3: Learning Curves on Reuters Data

The learning curves of SED, Convex TED and Margin are plotted in Figure 3. From the figure, SED outperforms Convex TED and Margin especially in the early stage. This observation again validates the contribution of label information to experimental design.

We also plot the learning curves of SED, Convex TED and Margin for the four tasks in Figure 4. SED again outperforms its counterpart, Convex TED, for all tasks. It is interesting to observe that when the data are rather balanced, such as in *MCAT*, Margin performs better than Convex TED. This is actually possible, because when the data are balanced, discriminative examples near the class boundary will provide the most useful information to the learner. Note that the effectiveness of SED can be further improved if we put more weight on the label information for this task. Nevertheless, we leave the issue of automatically learning the weight of label information to our future research.



(a) Autos (b) Motorcycles

(c) Baseball (d) Hockey

Figure 2: Learning Curves for Four Binary Classification Tasks on Newsgroups Data

Table 3: Comparison of Methods (in Average AUC) on Reuters Data

$\|\mathcal{L}\|$	SED	Convex TED	Sequential TED	Margin	Random Sampling
5	0.7347±0.0416	0.7347±0.0416	**0.7910**±0.0250	0.7347±0.0401	0.6932±0.0669
10	**0.8215**±0.0201	0.8039±0.0267	0.8137±0.0232	0.7942±0.0264	0.7579±0.0434
15	**0.8565**±0.0216	0.8327±0.0190	0.8397±0.0135	0.8341±0.0203	0.7945±0.0246
20	**0.8720**±0.0117	0.8416±0.0154	0.8551±0.0117	0.8538±0.0215	0.8235±0.0227
25	**0.8842**±0.0066	0.8550±0.0126	0.8668±0.0131	0.8646±0.0233	0.8383±0.0190
30	**0.8902**±0.0078	0.8645±0.0106	0.8725±0.0137	0.8775±0.0168	0.8536±0.0226
35	**0.8963**±0.0069	0.8709±0.0091	0.8807±0.0139	0.8897±0.0114	0.8684±0.0162
40	0.8992±0.0088	0.8774±0.0085	0.8850±0.0118	**0.8993**±0.0129	0.8749±0.0130
45	0.9024±0.0085	0.8789±0.0080	0.8898±0.0126	**0.9042**±0.0117	0.8864±0.0108
50	0.9048±0.0087	0.8856±0.0068	0.8932±0.0120	**0.9093**±0.0101	0.8948±0.0096

4.4 Discussions

4.4.1 *Effectiveness of Label Information*

As we have discussed in Section 3, the contribution of label information is controlled by the parameter γ_2. If $\gamma_2 = 0$, we do not use the label information at all and hence our method degenerates to Convex TED; as γ_2 increases, we put larger weight on the label information. To evaluate the contribution of label information, we carry out a set of experiments by varying the value of γ_2 in the *autos* task of the Newsgroups data set. As before, each candidate set covers 50% of the unlabeled documents and the parameters are set to be $\mu = 0.01, \gamma_1 = 0.1\gamma_{\max}$.

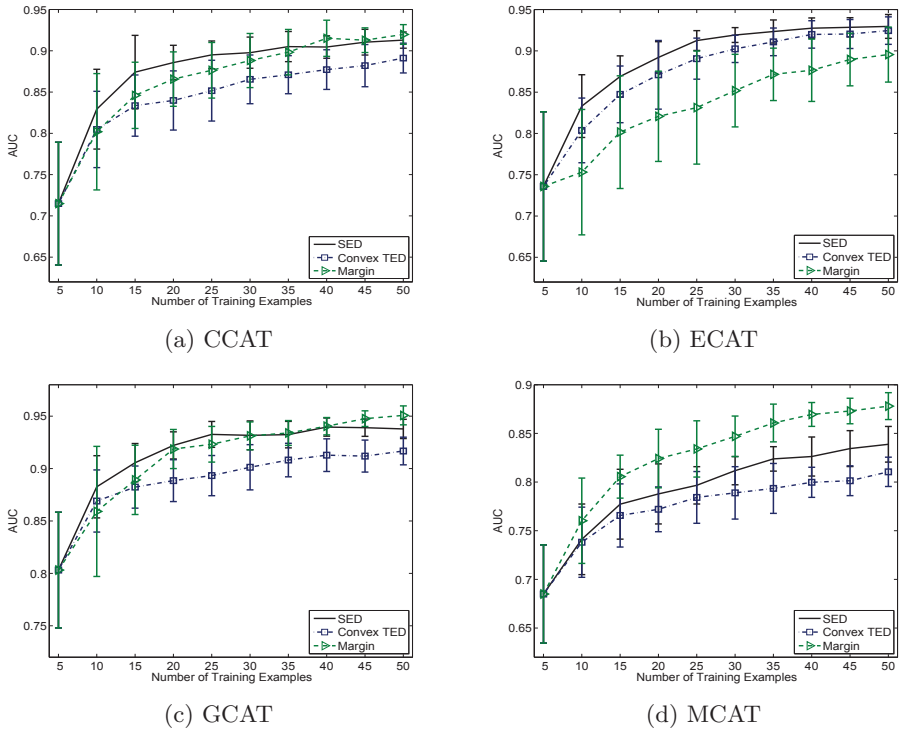
The learning curves of SED with different γ_2 values are plotted in Figure 5. As we can see, using a large enough γ_2 value, e.g. $\gamma_2 = 1$, can greatly speed up the learning procedure, while using small values, e.g. $\gamma_2 = [0, 0.1]$, will not improve much.

This observation validates the effectiveness of label information for experimental design. It should be noted that if γ_2 is too large, e.g. $\gamma_2 = 10$ or 100, the learning rate will be slower than that with moderate γ_2 values in the early stage of learning.

This is because the training set is too small in this stage and the class boundary learned is not very accurate, so adopting too large γ_2 values will mislead example selection by querying the outliers. The risk can be mitigated as the size of the training set increases. We also note that choosing $\gamma_2 = 1$ will achieve the best performance not only in the early stage but also in the later stage.

Similar experiments are conducted for the *MCAT* task of the Reuters data set. As before, the random candidate sets cover 20% of the unlabeled documents and the parameters are set to be $\mu = 0.01, \gamma_1 = 0.1\gamma_{\max}$. The learning curves of SED with different γ_2 values are plotted in Figure 6.

From Figure 6, it is interesting to observe that, different from what we have found in Figure 5, adopting a larger value of γ_2 will always improve the active learning procedure. This is because in the *MCAT* task, about 50% of the documents are positive, but in the *autos* task, only 25% of



(a) CCAT (b) ECAT

(c) GCAT (d) MCAT

Figure 4: Learning Curves for Four Binary Classification Tasks on Reuters Data

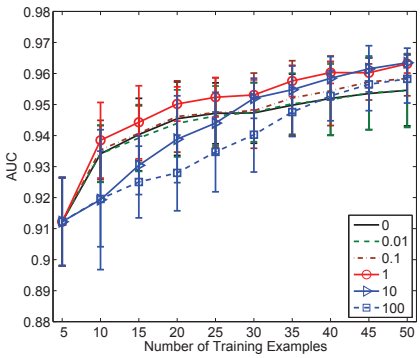


Figure 5: Effect of Varying γ_2 on *Autos* Task

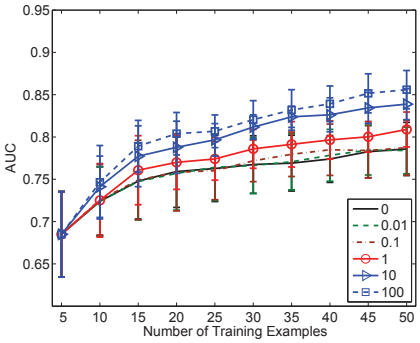


Figure 6: Effect of Varying γ_2 on *MCAT* Task

the documents are positive. Since the data distribution is more balanced in the *MCAT* task, adopting a larger value of γ_2 will always choose those discriminative examples without taking the risk of querying the outliers.

4.4.2 *Effect of Candidate Set Size*

In this section, we conduct several experiments to investigate the effect of the candidate set size by randomly choosing 20%, 40%, 60% and 80% of the unlabeled documents to form different candidate sets. For the *autos* task, the parameters are set to be $\mu = 0.01, \gamma_1 = 0.1\gamma_{\max}, \gamma_2 = 1$, and the learning curves for different candidate set sizes are plotted in Figure 7. For the *MCAT* task, the parameters are set to be $\mu = 0.01, \gamma_1 = 0.1\gamma_{\max}, \gamma_2 = 10$, and the learning curves are plotted in Figure 8.

As we can see from Figure 7, using a larger candidate set will greatly speed up the learning procedure. We note that as the size of the candidate set increases, the performance gap between the learned classifiers becomes smaller. However, in Figure 8, the learning curves are not so sensitive to the candidate set size as those in Figure 7. This can again be explained by the distribution of data. The more balanced the data are, the less sensitive is the method to the candidate set size. We should note that the candidate set size has great impact on the optimization problem. Specifically, the larger the candidate set is, the longer time we need to solve the problem. Thus, in practice, we should maintain a tradeoff between performance and the time cost and use a candidate set of a reasonable size.

5. CONCLUSION

In this paper, we have proposed a novel active learning method, SED, to seamlessly incorporate label information into the document selection procedure of experimental design. To the best of our knowledge, SED is the first work that uses label information to improve experimental design. One promising property of SED is that it can effectively use label information and the distribution of unlabeled data in a

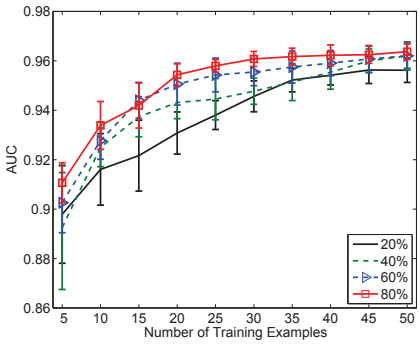


Figure 7: Effect of Candidate Set Size on *Autos* Task

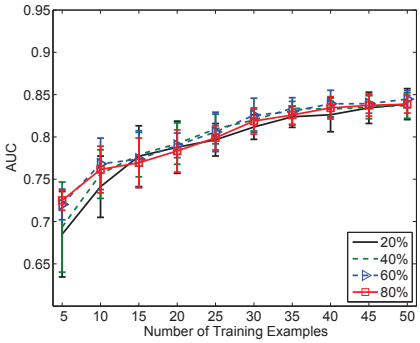


Figure 8: Effect of Candidate Set Size on *MCAT* Task

unified framework. In particular, SED can greatly speed up the learning procedure when the distribution of unlabeled data is balanced, while existing methods based on experimental design always perform badly in this case. Moreover, SED can greatly outperform margin-based active learning when the distribution of unlabeled data is unbalanced. As another promising property, SED is convex and thus global optimality is guaranteed. Experiments conducted on two text corpora demonstrate that SED outperforms state-of-the-art active learning algorithms, such as TED and margin-based methods, which take into consideration only partial information.

One of our future research directions is to automatically learn from data the contribution of label information, i.e. γ_2. Another possible research direction is to apply SED to other information retrieval applications.

6. ACKNOWLEDGMENTS

The authors thank Wu-Jun Li and Ning Zhu for some helpful discussions. This research has been supported by General Research Fund 622209 from the Research Grants Council of Hong Kong.

7. REFERENCES

[1] D. Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1988.

[2] A. Atkinson and A. Donev. *Optimum Experimental Designs*. Oxford University Press, USA, 1992.

[3] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Mach. Learn.*, 15(2):201–221, 1994.

[4] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *J. Artif. Intell. Res.*, 4:129–145, 1995.

[5] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, 2003.

[6] Y. Guo and R. Greiner. Optimistic active learning using mutual information. In *IJCAI*, 2007.

[7] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *NIPS*, 2007.

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

[9] S. C. Hoi, R. Jin, and M. R. Lyu. Large-scale text categorization by batch mode active learning. In *WWW*, 2006.

[10] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *ICML*, 2006.

[11] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, 1994.

[12] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

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

[14] A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. In *ICML*, 1998.

[15] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML*, 2004.

[16] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[17] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT*, 1992.

[18] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, 2002.

[19] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR*, 1999.

[20] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *ICML*, 2006.

[21] K. Yu, S. Zhu, W. Xu, and Y. Gong. Non-greedy active learning for text categorization using convex transductive experimental design. In *SIGIR*, 2008.

[22] J. Zhang and Y. Yang. Robustness of regularized linear classification methods in text categorization. In *SIGIR*, 2003.

[23] L. Zhang, C. Chen, W. Chen, J. Bu, D. Cai, and X. He. Convex experimental design using manifold structure for image retrieval. In *ACM MM*, 2009.

[24] T. Zhang and F. J. Oles. Text categorization based on regularized linear classification methods. *Inform. Retrieval*, 4(1):5–31, 2001.

[25] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML Workshop*, 2003.