Early Active Learning via Robust Representation and Structured Sparsity

Feiping Nie[†], Hua Wang[‡], Heng Huang[†*], Chris Ding[†]
[†]Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
[‡]Department of Electrical Engineering and Computer Science
Colorado School of Mines, Golden, Colorado 80401, USA
feipingnie@gmail.com, huawangcs@gmail.com, heng@uta.edu, chqding@uta.edu

Abstract

Labeling training data is quite time-consuming but essential for supervised learning models. To solve this problem, the active learning has been studied and applied to select the informative and representative data points for labeling. However, during the early stage of experiments, only a small number (or none) of labeled data points exist, thus the most representative samples should be selected first. In this paper, we propose a novel robust active learning method to handle the early stage experimental design problem and select the most representative data points. Selecting the representative samples is an NP-hard problem, thus we employ the structured sparsity-inducing norm to relax the objective to an efficient convex formulation. Meanwhile, the robust sparse representation loss function is utilized to reduce the effect of outliers. A new efficient optimization algorithm is introduced to solve our non-smooth objective with low computational cost and proved global convergence. Empirical results on both single-label and multi-label classification benchmark data sets show the promising results of our method.

1 Introduction

In many machine learning tasks, data and label collections are time-consuming and costly. Thus, it is desired to find ways to minimize the number of data points to be labeled, *i.e.* carefully select the data points that should be labeled by users. Recent research on active learning has been reasonably successful in handling the problem of selecting examples to be labeled. Typically, the active learning algorithms choose which data points should be labeled and added to the training set. In other words, a learner begins with a small set of labeled data, selects a few informative data from the unlabeled data, and queries for labels from an oracle. The goal is to reduce the total labeling cost incurred to train an accurate supervised learning model.

*Corresponding Author. This work was partially supported by NSF CCF-0830780, CCF-0917274, DMS-0915228, IIS-1117965.

In most cases, the active learning methods build upon notions of uncertainty in classification. For example, the informative data points that are most likely to be misclassified should be considered to be the most informative and will be selected for supervision; on the other hand, the representative data points that capture the most important natural properties of data points (*e.g.* manifold, clustering structures, *etc.*) should also be selected. Starting from these two different statistical uncertainty point of views, the existing active learning approaches can be categorized into two groups. The first group methods query the most informative data points, such as uncertainty sampling [Lewis and Catlett, 1994; Balcan *et al.*, 2007], query by committee [Seung *et al.*, 1992; Freund *et al.*, 1997], and optimal experimental design [Lindley, 1956; Flaherty *et al.*, 2005]. The other group methods target to select the data points with representative information, such as transductive experimental design [Yu *et al.*, 2006] and clustering based method [Nguyen and Smeulders, 2004; Nie *et al.*, 2012].

The active learning algorithms in both categories solve the experimental design problem at different stages. The first category methods prefer to choosing the uncertain or hard to be predicted data, but need determine them by a large number of labeled samples to avoid the samples bias. Thus, such methods should be used during the mid-stage of experiments, *i.e.* after enough labeled samples are collected. At the early stage of experiments, because we have a small number (or even none) of labeled data, the representative data points are desired to be selected for supervision.

In this paper, we focus on the early active learning strategies, *i.e.* solving the early stage experimental design problem. The Transductive Experimental Design (TED) method was proposed to select the data points via a least square loss function and ridge regularization [Yu *et al.*, 2006]. However, the TED objective leads to an NP-hard problem, and was approximated by a sequential optimization problem with slow greedy algorithm [Yu *et al.*, 2006]. Meanwhile, the least square loss function used in TED objective is sensitive to data outliers, thus the TED method is not suitable for the real-world applications with large noise.

To solve these two deficiencies, we propose a novel robust active learning approach using the structured sparsity-inducing norms to relax the NP-hard objective to the convex formulation. A new robust active learning loss function is in-

troduced to selected the representative data insensitive to the effect of outliers. The structured sparse regularization is utilized to select the representative data points that have non-zero weights during the sparse representations of different data (the non-representative data points typically have small weights or even zero, when they are used to represent other samples). A new optimization method is derived to solve our non-smooth objective with efficient procedure. Our algorithm has the closed form solution in each iteration with proved global convergence. Compared to previous greedy algorithm [Yu *et al.*, 2006], our new optimization method is much more efficient. We evaluate our method using several benchmark data sets. All experimental results show our method outperforms other state-of-the-art active learning methods.

Notations. In this paper, matrices are written as boldface uppercase letters and vectors are written as boldface lowercase letters. Given a matrix $\mathbf{M} = \{m_{ij}\}$, we denote its i-th row, j-th column as $\mathbf{m}^i$, $\mathbf{m}_j$. The ℓ_2-norm of a vector $\mathbf{v}$ is defined as $||\mathbf{v}||_2 = \sqrt{\mathbf{v}^T\mathbf{v}}$. The $\ell_{2,1}$-norm of a matrix is defined as $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n}\left\|\mathbf{m}^i\right\|_2$. For consistency, the quasi-norm $\ell_{2,0}$-norm of a matrix $\mathbf{M}$ is defined as the number of the nonzero rows of $\mathbf{M}$.

2 Active Learning via Structured Sparsity-Inducing Norms

Given unlabeled data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, the task of active learning is to select m $(m < n)$ samples for labeling by user, such that the potential performance is maximized when training with the fixed m labeled samples. In early stage experimental design, the early active learning methods usually select the representative data points to maximize the efficiency of training process. Such a problem can be formulated as the Transductive Experimental Design (TED) problem [Yu *et al.*, 2006] to solve:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{B}\mathbf{a}_i\|_2^2 + \gamma\|\mathbf{a}_i\|_2^2\right) \\ &s.t.\ \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m\times n},\ \ \mathbf{B} \subset \mathbf{X}, |\mathbf{B}| = m. \end{aligned} \tag{1}$$

The key idea of TED is to select the samples that can best represent the whole data using a linear representation. Therefore, the selected samples can be considered to be the most representative and informative data points since the other data points are highly linearly related to the selected samples.

However, Eq. (1) leads to a difficult combinatorial optimization problem (NP-hard). Thus, the TED problem was approximated by a sequential optimization problem and solved by an inefficient greedy optimization approach. On the other hand, the TED objective minimizes the least square error, hence it is sensitive to the data outliers. To solve these two deficiencies, we formulate the early active learning problem using the structured sparsity-inducing norms and propose a new robust formulation with efficient optimization algorithm.

Using the $\ell_{2,0}$-norm, we formulate the problem in Eq. (1) as:

$$\begin{aligned} &\min_{\mathbf{A}} \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \gamma\|\mathbf{a}_i\|_2^2\right) \\ &s.t.\quad \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{n\times n},\ \ \|\mathbf{A}\|_{2,0} = m. \end{aligned} \tag{2}$$

We propose a more concise objective as following:

$$\min_{\mathbf{A}} \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \gamma\|\mathbf{A}\|_{2,0}\,. \tag{3}$$

However, solving this problem is NP-hard because of the $\ell_{2,0}$-norm. Recent theoretical progress shows that $\|\mathbf{A}\|_{2,1}$ is the minimum convex hull of $\|\mathbf{A}\|_{2,0}$, and when $\mathbf{A}$ is row-sparse enough, one can always minimize $\|\mathbf{A}\|_{2,1}$ to obtain the same result of minimizing $\|\mathbf{A}\|_{2,0}$. Therefore, Eq. (3) can be relaxed to the following convex optimization problem[1]:

$$\min_{\mathbf{A}} \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \gamma\|\mathbf{A}\|_{2,1}\,. \tag{4}$$

The first term in Eq. (4) is the loss of the representation, and the second term in Eq. (4) is the $\ell_{2,1}$-norm regularization on $\mathbf{A}$ to obtain row-sparse solution of $\mathbf{A}$. The loss used in Eq. (4) is a squared loss, which is very sensitive to data outliers, thus we propose to solve the following problem:

$$\min_{\mathbf{A}} \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2 + \gamma\|\mathbf{A}\|_{2,1}\,, \tag{5}$$

which can be written into a matrix format:

$$J = \min_{\mathbf{A}} \left\|(\mathbf{X} - \mathbf{X}\mathbf{A})^T\right\|_{2,1} + \gamma\|\mathbf{A}\|_{2,1}\,. \tag{6}$$

In our new objective, where the $\ell_{2,1}$-norm is applied to the loss function, the ℓ_1-norm is imposed among data points and the ℓ_2-norm is used for features. As a result, the effects of data outliers are reduced and the robustness of active learning is improved.

After obtaining the optimal $\mathbf{A}$ in Eq. (6), we can sort the rows of $\mathbf{A}$ by the row-sum values of the absolute $\mathbf{A}$ in the decreasing order. Therefore, the active learning task can be performed by selecting the m samples corresponding to the top m rows of $\mathbf{A}$. As the proposed active learning method is based on robust representation and structured sparsity, we call our active learning method as RRSS for short.

3 Optimization Algorithm and Analysis

Although the problem (6) is convex, it is difficult to be solved since two non-smooth terms are involved. In this section, we will derive an efficient algorithm to solve the problem (6), and give a theoretical analysis to show that the algorithm will converge to the globally optimal solution.

3.1 Algorithm Derivation

Taking the derivative of Eq. (6) w.r.t $\mathbf{A}$, and setting the derivative to zero, we have[2]:

$$\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{U} - \mathbf{X}^T\mathbf{X}\mathbf{U} + \gamma\mathbf{V}\mathbf{A} = \mathbf{0} \tag{7}$$

[1] Recently, we observed an equivalent formulation of Eq.(4) in [Yu *et al.*, 2008] with different motivation.

[2] When $\mathbf{x}_i - \mathbf{X}\mathbf{a}_i = \mathbf{0}$, we can regularize u_{ii} as $u_{ii} = \frac{1}{2\sqrt{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \varsigma}}$. Similarly, when $\mathbf{a}^i = \mathbf{0}$, $v_{ii} = \frac{1}{2\sqrt{\|\mathbf{a}^i\|_2^2 + \varsigma}}$. Then the derived algorithm can be proved to minimize $\sum_{i=1}^{n}\sqrt{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \varsigma} + \gamma\sum_{i=1}^{n}\sqrt{\|\mathbf{a}^i\|_2^2 + \varsigma}$. It is easy to see that the problem is reduced to problem (6) when $\varsigma \to 0$.

where $\mathbf{U}$ is a diagonal matrix with the i-th diagonal element as $u_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$, and $\mathbf{V}$ is a diagonal matrix with the i-th diagonal element as $v_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. Then for each $i(1 \leq i \leq n)$, we have

$$u_{ii}\mathbf{X}^T\mathbf{X}\mathbf{a}_i - u_{ii}\mathbf{X}^T\mathbf{x}_i + \gamma\mathbf{V}\mathbf{a}_i = \mathbf{0} \tag{8}$$

and then $\mathbf{a}_i$ can be calculated by

$$\mathbf{a}_i = u_{ii}(u_{ii}\mathbf{X}^T\mathbf{X} + \gamma\mathbf{V})^{-1}\mathbf{X}^T\mathbf{x}_i \tag{9}$$

Note that $\mathbf{U}$ and $\mathbf{V}$ both are dependent on $\mathbf{A}$ and thus is also unknown variables, we propose an iterative algorithm to solve this problem. The detailed algorithm is described in Algorithm 1. In the following, we will prove that the algorithm will converge to the global optimal solution to the problem (6).

Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$.
Initialize $\mathbf{A} \in \mathbb{R}^{n\times n}$;
while *not converge* **do**
  1. Calculate the diagonal matrix $\mathbf{U}$, where the i-th diagonal element of $\mathbf{U}$ is $u_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$.
  Calculate the diagonal matrix $\mathbf{V}$, where the i-th diagonal element of $\mathbf{V}$ is $v_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. ;
  2. For each $i(1 \leq i \leq n)$, update $\mathbf{a}_i$ by $\mathbf{a}_i = u_{ii}(u_{ii}\mathbf{X}^T\mathbf{X} + \gamma\mathbf{V})^{-1}\mathbf{X}^T\mathbf{x}_i$;
end
Output: The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Algorithm 1: The algorithm to solve the problem (6).

3.2 Algorithm Convergence Analysis

First, we introduce a lemma as follows:

Lemma 1 ([Nie *et al.*, 2010]) *For any vectors* $\tilde{\mathbf{a}}$ *and* $\mathbf{a}$*, we have* $\|\tilde{\mathbf{a}}\|_2 - \frac{\|\tilde{\mathbf{a}}\|_2^2}{2\|\mathbf{a}\|_2} \leq \|\mathbf{a}\|_2 - \frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{a}\|_2}$.

Next, we prove the convergence of our algorithm in the following theorem:

Theorem 1 *Algorithm 1 decreases the objective value in each iteration.*

Proof: Suppose the updated $\mathbf{A}$ is $\tilde{\mathbf{A}}$. According to the step 2 in Algorithm 1, we know that

$$\begin{aligned}\tilde{\mathbf{A}} = \ & \arg\max_{\mathbf{M}} Tr((\mathbf{X} - \mathbf{X}\mathbf{M})\mathbf{U}(\mathbf{X} - \mathbf{X}\mathbf{M})^T) \\ & + \gamma Tr(\mathbf{M}^T\mathbf{V}\mathbf{M})\end{aligned} \tag{10}$$

Thus we have

$$\begin{aligned} & Tr((\mathbf{X} - \mathbf{X}\tilde{\mathbf{A}})\mathbf{U}(\mathbf{X} - \mathbf{X}\tilde{\mathbf{A}})^T) + \gamma Tr(\tilde{\mathbf{A}}^T\mathbf{V}\tilde{\mathbf{A}}) \\ \leq \ & Tr((\mathbf{X} - \mathbf{X}\mathbf{A})\mathbf{U}(\mathbf{X} - \mathbf{X}\mathbf{A})^T) + \gamma Tr(\mathbf{A}^T\mathbf{V}\mathbf{A})\end{aligned}$$

According to the definitions of $\mathbf{U}$ and $\mathbf{V}$, we have

$$\begin{aligned} & \sum_{i=1}^{n}\left(\frac{\|\mathbf{x}_i - \mathbf{X}\tilde{\mathbf{a}}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2} + \gamma\frac{\|\tilde{\mathbf{a}}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right) \\ \leq \ & \sum_{i=1}^{n}\left(\frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2} + \gamma\frac{\|\mathbf{a}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right)\end{aligned} \tag{11}$$

According to the Lemma 1, we have the following two inequalities:

$$\begin{aligned} & \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\tilde{\mathbf{a}}_i\|_2 - \frac{\|\mathbf{x}_i - \mathbf{X}\tilde{\mathbf{a}}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}\right) \\ \leq \ & \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2 - \frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}\right)\end{aligned} \tag{12}$$

$$\gamma\sum_{i=1}^{n}\left(\|\tilde{\mathbf{a}}^i\|_2 - \frac{\|\tilde{\mathbf{a}}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right) \leq \gamma\sum_{i=1}^{n}\left(\|\mathbf{a}^i\|_2 - \frac{\|\mathbf{a}^i\|_2^2}{2\|\mathbf{a}^i\|_2}\right) \tag{13}$$

Summing Eq. (11)-(13) in the two sides, we arrive at

$$\sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\tilde{\mathbf{a}}_i\|_2 + \gamma\|\tilde{\mathbf{a}}^i\|_2\right) \leq \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2 + \gamma\|\mathbf{a}^i\|_2\right) \tag{14}$$

Therefore, the algorithm decreases the objective value in each iteration. □

Denote $f(\mathbf{M}) = Tr((\mathbf{X} - \mathbf{X}\mathbf{M})\mathbf{U}(\mathbf{X} - \mathbf{X}\mathbf{M})^T) + \gamma Tr(\mathbf{M}^T\mathbf{V}\mathbf{M})$. In the convergence, we know $f(\tilde{\mathbf{A}}) = f(\mathbf{A})$. Note that $\tilde{\mathbf{A}} = \arg\max_{\mathbf{M}} f(\mathbf{M})$, thus $\mathbf{A} = \arg\max_{\mathbf{M}} f(\mathbf{M})$. Note that $\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{0}$ is exactly the Eq. (7), thus $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{V}$ will satisfy the Eq. (7) in the convergence. As the problem (6) is a convex problem, satisfying the Eq. (7) indicates that $\mathbf{A}$ is a global optimum solution to the problem (6). Therefore, the Algorithm 1 will converge to the global optimum of the problem (6). Because we have closed form solution in each iteration, our algorithm converges very fast.

4 Kernel Extension

The proposed method in the previous section can be easily extended to the kernel version for active learning or we can simply use a general kernelization framework [Zhang *et al.*, 2010] to achieve the kernel version. Let $\phi : \mathbb{R}^d \to \mathcal{H}$ be a kernel mapping from the original space to the kernel space, where $\mathcal{H}$ is a reproducing Kernel Hilbert Space (RKHS) induced by a kernel function $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T\phi(\mathbf{z})$. Then $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_n)]$, and the problem (6) in the kernel space becomes:

$$\min_{\mathbf{A}} \|(\phi(\mathbf{X}) - \phi(\mathbf{X})\mathbf{A})^T\|_{2,1} + \gamma\|\mathbf{A}\|_{2,1} \tag{15}$$

Given the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, we denote the kernel matrix by $\mathbf{K} \in \mathbb{R}^{n\times n}$, where the (i, j)-th element of $\mathbf{K}$ is $k_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. The algorithm to solve the problem (15) is very similar to the Algorithm 1, and the detailed algorithm is described in Algorithm 2.

5 Experimental Results

We empirically evaluate the proposed RRSS method, as well as its kernelized version, K-RRSS method. We first demonstrate the effectiveness of the proposed methods on a challenging synthetic data set. Then we study their performances in both single-label and multi-label classification tasks.

Input: The kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$.
Initialize $\mathbf{A} \in \mathbb{R}^{n \times n}$;
while *not converge* **do**
1. Calculate the diagonal matrix $\mathbf{U}$, where the i-th diagonal element of $\mathbf{U}$ is $u_{ii} = \frac{2}{\sqrt{k_{ii} - 2\mathbf{a}_i^T \mathbf{k}_i + \mathbf{a}_i^T \mathbf{K} \mathbf{a}_i}}$. Calculate the diagonal matrix $\mathbf{V}$, where the i-th diagonal element of $\mathbf{V}$ is $v_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. ;
2. For each $i (1 \leq i \leq c)$, update $\mathbf{a}_i$ by $\mathbf{a}_i = u_{ii}(u_{ii}\mathbf{K} + \gamma \mathbf{V})^{-1}\mathbf{k}_i$;
end
Output: The matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Algorithm 2: The algorithm to solve the problem (15).

5.1 Experiment on A Challenging Synthetic Data

We first give an experimental result on a challenging synthetic data to validate the effectiveness of the proposed active learning algorithm. The synthetic data are generated from six subspaces of the original 500-dimension space. The dimensions of the six subspaces are 10, 12, 14, 16, 18, 20, respectively. The numbers of data generated from the six subspaces are 25, 30, 35, 40, 45, 50, respectively, which are relatively small to the dimensions. Furthermore, we add noise of 20% perturbation on the generated subspaces data. Thus the 500-dimensional data are distributed on six subspaces with different dimensions, different numbers and high noises. Separating such a high dimensional, high noised and small sampled data from the six subspaces is a very challenging task, which can be seen from the 10-nearest-neighbor similarity graph of the data in Figure 1(a). The cluster structure is obscure and can not be discerned from the similarity graph on the original data. Figure 1(b) shows the learned matrix A in Eq. (6) by Algorithm 1, and Figure 1(c) shows the selected top 20 rows of A, the numbers of the selected samples from the six subspaces are 1, 2, 3, 3, 4, 7. The result clearly indicates the effectiveness of the proposed active learning algorithm.

5.2 Improved Single-Label Classification

We evaluate the proposed methods in single-label classification tasks, in which each data point belongs to one and only one class. We experiment with the following four broadly used single-label data sets: Yale face image data [Georghiades *et al.*, 2001], USPS, BinAlpha [Belhumeur *et al.*, 1997], and Coil20 [Nene *et al.*, 1996].

Yale face database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration.

USPS database is a handwritten digit database. We randomly pick up 2000 images, 200 for each digit, from the database for our experiments.

BinAlpha data set contains 26 binary hand-written alphabets and we randomly select 30 images for every alphabet.

Coil20 database contains 20 objects. The images of each objects were taken 5 degrees apart as the object is rotated on a turntable and each object has 72 images. The size of each image is 32×32 pixels, with 256 grey levels per pixel.

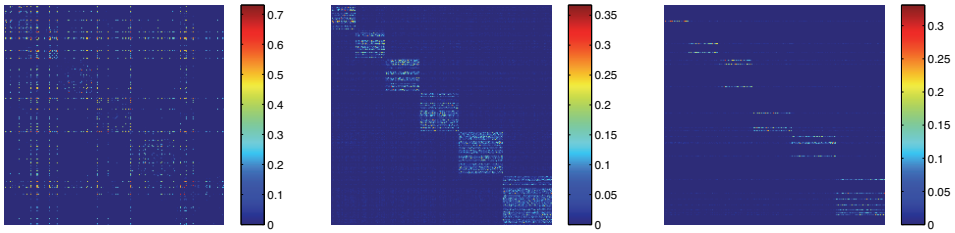
Experimental setups. We conduct our experiments as following. For each data set, we first randomly select 50% of the data points as candidate samples for training, from which we employ the compared active learning methods to select a certain number of samples to request human labeling. Using the selected samples and their queried labels as training data, we learn a classification model, by which we classify the other 50% data points. The latter is considered as test data. We repeat every test case for 50 times and report the average classification performances.

We compare our methods against the following two baseline methods including: (1) random sample selection method that randomly selects data points from a given data set to query labels; (2) K-means sample selection method, where we perform K-means clustering and select the data points that are most close to the K centroids. The former is equivalent to traditional passive learning, while the latter is to seek the most representative samples of a data set. We also compare our method against two related active learning methods including (3) transductive experimental design (TED) method [Yu *et al.*, 2006] and (4) QUIRE method [Huang *et al.*, 2010]. The former is closely related to our method, and the latter is a very recent method that has demonstrated state-of-the-art performance.

We construct a Gaussian kernel from the candidate data points of each data set, *i.e.*, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\alpha \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)$, as input for TED method, QUIRE method, and the proposed K-RRSS method. Prior to experiments, for each data set we use support vector machine (SVM) (linear kernel) to seek the optimal parameter α of the Gaussian kernel and the optimal tradeoff parameter of the three methods via a standard 5-fold cross-validation using the candidate data in the range of $\left\{10^{-5}, \ldots, 1, \ldots, 10^5\right\}$, where we select 50% of the training data of each of the 5 trails to learn the SVM classifier.

We learn SVM and transductive SVM [Joachims, 1999] classifiers, by which we classify the test data points. The former is one of the most widely used supervised single-label classification method, while the latter is an established semi-supervised method extended from the former. Again, Gaussian kernel is used, where we set $C = 1$ and use the fine tuned γ in the sample selection process.

Experimental results. The classification accuracies by SVMs trained with different amount of samples selected by the compared methods for each of the four data sets are shown in Figure 2. A first glance at the results shows that the proposed methods consistently outperform the compared methods, sometimes very significantly, which demonstrate their effectiveness in selecting useful training samples in single-label classification. By a more careful observation, we can see that the proposed K-RRSS method is always better than its counterpart using vector input (equivalent to using linear kernel), *i.e.*, RRSS method, which is consistent with their mathematical formulations in that kernel could make the data more discriminative in the mapped feature spaces. Moreover, the performance curves of the two proposed methods climb to saturation in a much faster way than the compared methods. That is, when the number selected samples is small, our



(a) 10-NN similarity graph (b) Learned matrix A (c) The top 20 rows of A

Figure 1: The experimental result on the noised subspace data. Left figure: the 10-nearest-neighbor similarity graph on the original data. Middle figure: the learned matrix A by Algorithm 1. Right figure: the selected top 20 rows of A, the numbers of the selected samples from the six subspaces are 1, 2, 3, 3, 4, 7.

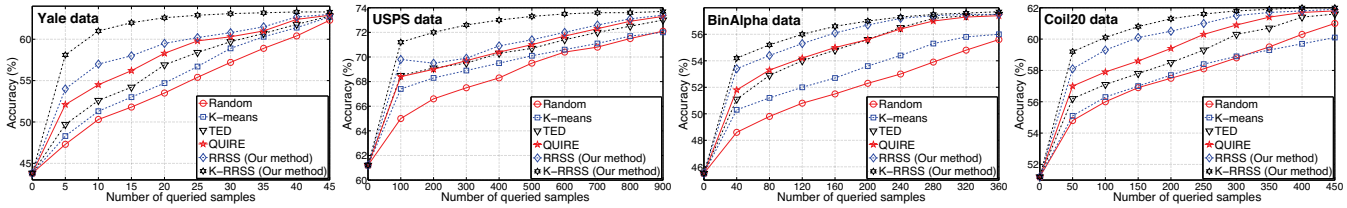


Figure 2: Classification accuracies using SVMs trained by varied amount of samples selected by the active learning methods.

Table 1: Classification accuracies using transductive SVMs trained by the top 10% of the candidate images selected by the compared active learning methods.

	Yale	USPS	BinAlpha	Coil20
Random	0.523	0.667	0.511	0.572
K-means	0.537	0.673	0.519	0.576
TED	0.558	0.683	0.528	0.579
QUIRE	0.569	0.688	0.532	0.585
RRSS	0.588	0.710	0.537	0.596
K-RRSS	**0.631**	**0.723**	**0.545**	**0.612**

methods are more effective in selecting the most useful data points to train an classifier with high accuracy. Because human annotation is often expensive in real world applications, this adds to the practical value of the proposed methods. Finally, the performances of the active learning methods including ours are generally superior to the passive learning method that randomly selects samples for label query, which confirms the values of active learning.

Table 1 show the classification results using transductive SVMs trained with the top 10% selected samples by the compared methods on each data set. We conduct this experiment because semi-supervised learning is another way to improve the classification performance by exploiting unlabeled data, which could work with active learning and reinforce each other. Again, our methods demonstrate better performances, which confirms their effectiveness from another perspective.

5.3 Improved Multi-Label Classification

Now we evaluate the proposed approach in multi-label classification tasks, in which one single data point could be associated with more than one class label. Multi-label classification is more general, though more challenging, than single-label classification, therefore it is more close to real world applications. One of the most successful application of multi-label classification is image annotation, because one picture usually contains more than one object of interest. We evaluate the proposed methods on the following three multi-label benchmark image data sets: TRECVID 2005, MSRC, and PASCAL VOC 2010.

TRECVID 2005 data set contains 137 broadcast videos (74523 sub-shot) from 13 different programs. Following previous work [Wang *et al.*, 2009], we randomly sample the data such that each concept has at least 100 images.

MSRC data set contains 591 images with 23 classes. Around 80% of the images are annotated with at least one classes and around three classes per images on average.

PASCAL VOC 2010 data set [Everingham *et al.*,] contains 13321 images with 20 classes. We randomly select images, such that at least 100 images are selected for each class, which leads to 1864 images used in our experiments.

Following prior computer vision studies, for these three image data sets, we divide each image into 64 blocks by a 8×8 grid and compute the first and second moments (mean and variance) of each color band to obtain a 384-dimensional vector as features.

Experimental setups. We follow the same experimental setups as in single label classification, *i.e.*, we randomly split a data set into two parts with equal size, and use one half as candidate pool and the other as test data. We still use the six compared methods as before to select images with the same settings. Upon the top 20% selected images and their queried labels, we classify the test images using the multi-label k-Nearest Neighbor (Mk-NN) method [Zhang and Zhou, 2007], which is a lazy learning method without a training process.

Multi-label classification is more complicated than single-

Table 2: Classification performance of Mk-NN method using the top 20% selected training samples by compared active learning methods on multi-label data sets.

	Methods	Hamming loss ↓	One-error ↓	Coverage ↓	Rank loss ↓	Average precision ↑
TRECVID	Random	0.187	0.368	2.642	0.234	0.467
	Kmeans	0.181	0.360	2.635	0.228	0.468
	TED	0.176	0.342	2.621	0.226	0.471
	QUIRE	0.168	0.336	2.611	0.218	0.477
	RRSS	0.148	0.319	2.584	0.201	0.488
	K-RRSS	**0.135**	**0.306**	**2.473**	**0.189**	**0.497**
MSRC	Random	0.289	0.554	3.969	0.293	0.571
	Kmeans	0.281	0.550	3.884	0.288	0.579
	TED	0.268	0.537	3.712	0.271	0.586
	QUIRE	0.263	0.532	3.684	0.263	0.591
	RRSS	0.256	0.515	3.516	0.258	0.609
	K-RRSS	**0.250**	**0.511**	**3.507**	**0.248**	**0.611**
PASCAL	Random	0.182	0.312	1.069	0.196	0.439
	Kmeans	0.177	0.308	1.030	0.190	0.441
	TED	0.175	0.304	1.017	0.187	0.443
	QUIRE	0.172	0.301	0.996	0.181	0.449
	RRSS	0.171	0.297	0.994	0.176	0.453
	K-RRSS	**0.163**	**0.291**	**0.986**	**0.170**	**0.462**

label classification, therefore it requires more metrics to evaluate the classification performance. We employ the following five widely used multi-label performance metrics to assess the classification results: Hamming loss, one-error, coverage, rank loss and average precision. We refer readers to [Zhang and Zhou, 2007] for detailed definitions of these performance metrics. For the first four metrics, the smaller is better (denoted as ↓ in Table 2); while for the last one, the bigger is better (denoted as ↑ in Table 2).

Experimental results. The average classification performances on the three multi-label image data sets over 50 repeats are reported in Table 2. The results show that the proposed methods are again better than the compared methods, which demonstrate the effectiveness of our methods in selecting training images for multi-label data.

We further examine the average number of labels associated with the selected training images by the compared methods, which are listed in Table 3. Compared to the average numbers of labels associated with the images in the original data sets, the randomly selected images have about the same numbers of labels as the original data, while the images selected by the active learning methods, including ours, have greater average label numbers than the original data. Because the more labels a data point associated with in multi-label settings, the more information it conveys [Wang *et al.*, 2010], the results in Table 2 provide a concrete evidence that sample selection by active learning methods do improve the informativeness of the data. Because the average number of labels associated with the selected training images by our K-RRSS method is greatest, it is the most effective method in selecting content rich images for multi-label image classification.

Table 3: Average labels per image of the three data sets, and average labels per image of the selected training images by compared methods from these data sets.

	TRECVID 2005	MSRC	PASCAL
Original data	4.16	2.51	1.62
Random selected	4.09	2.56	1.64
K-means	4.35	2.66	1.69
TED selected	4.63	2.79	1.86
QUIRE	5.04	2.93	1.96
RRSS selected	5.87	4.07	2.03
K-RRSS selected	5.93	4.12	2.11

6 Conclusion

In this paper, we proposed a novel active learning method to solve the early stage experimental design problem. Instead of using the traditional least square loss function, we introduce the robust sparse representation based active learning loss function. As a result, the data points selected by our method are insensitive to the outliers. The $l_{2,1}$-norm based structured sparse regularization is utilized to select the most representative data points that have large weights in the sparse representations of other data points. We performed the empirical studies on both single-label and multi-label classification tasks. In all experimental results, our new approach outperforms other related active learning methods.

References

[Balcan *et al.*, 2007] M. F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *the 20th Annual Conference on Learning Theory*, pages 35–50, 2007.

[Belhumeur *et al.*, 1997] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on PAMI*, 19(7):711–720, July 1997.

[Everingham *et al.*,] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/.

[Flaherty *et al.*, 2005] P. Flaherty, M. I. Jordan, and A. P. Arkin. Robust design of biological experiments. In *Advances in Neural Information Processing Systems*, pages 363–370, 2005.

[Freund *et al.*, 1997] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. pages 133–168, 1997.

[Georghiades *et al.*, 2001] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on PAMI*, 23(6):643–660, 2001.

[Huang *et al.*, 2010] S.J. Huang, R. Jin, and Z.H. Zhou. Active Learning by Querying Informative and Representative Examples. In *NIPS*, 2010.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

[Lewis and Catlett, 1994] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *International Conference on Machine Learning*, pages 148–156, 1994.

[Lindley, 1956] D.V. Lindley. On the measure of information provided by an experiment. 27(4):986–1005, 1956.

[Nene *et al.*, 1996] S. A. Nene, S. K. Nayar, and H. Murase. *Columbia object image library (COIL-20), Technical Report CUCS-005-96*. Columbia University, 1996.

[Nguyen and Smeulders, 2004] H. T. Nguyen and A. W. M. Smeulders. Active learning using pre-clustering. In *International Conference on Machine Learning*, pages 623–630, 2004.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, 2010.

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

[Seung *et al.*, 1992] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *ACM Workshop on Computational Learning Theory*, pages 287–294, 1992.

[Wang *et al.*, 2009] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated Green's function. In *ICCV*, 2009.

[Wang *et al.*, 2010] H. Wang, C. Ding, and H. Huang. Multi-label Linear Discriminant Analysis. In *ECCV*, 2010.

[Yu *et al.*, 2006] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *ICML*, pages 1081–1088, 2006.

[Yu *et al.*, 2008] Kai Yu, Shenghuo Zhu, Wei Xu, and Yihong Gong. Non-greedy active learning for text categorization using convex ansductive experimental design. In *SIGIR*, pages 635–642, 2008.

[Zhang and Zhou, 2007] M.L. Zhang and Z.H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[Zhang *et al.*, 2010] Changshui Zhang, Feiping Nie, and Shiming Xiang. A general kernelization framework for learning algorithms based on kernel PCA. *Neurocomputing*, 73(4-6):959–967, 2010.