Active Learning via Neighborhood Reconstruction

Yao Hu Debing Zhang Zhongming Jin Deng Cai Xiaofei He
State Key Lab of CAD&CG, College of Computer Science,
Zhejiang University, Hangzhou 310058, China.
{huyao001,debingzhangchina,zhongmingjin888,dengcai,xiaofeihe}@gmail.com

Abstract

In many real world scenarios, active learning methods are used to select the most informative points for labeling to reduce the expensive human action. One direction for active learning is selecting the most representative points, ie., selecting the points that other points can be approximated by linear combination of the selected points. However, these methods fails to consider the local geometrical information of the data space. In this paper, we propose a novel framework named *Active Learning via Neighborhood Reconstruction* (ALNR) by taking into account the locality information directly during the selection. Specifically, for the linear reconstruction of target point, the nearer neighbors should have a greater effect and the selected points distant from the target point should be penalized severely. We further develop an efficient two-stage iterative procedure to solve the final optimization problem. Our empirical study shows encouraging results of the proposed algorithms in comparison to other state-of-the-art active learning algorithms on both synthetic and real visual data sets.

1 Introduction

In many applications, expensive human actions are required to collect label information. To reduce the cost of labeling, active learning methods are designed to choose the most informative examples (i.e., improve the classifier the most) to label for training, which have been shown to benefit many real world applications such as image retrieval [Gosselin and Cord, 2008], image and video classification [Yan *et al.*, 2003; Qi *et al.*, 2008], object categorization [Kapoor *et al.*, 2010] and document summarization [He *et al.*, 2012] and so on.

There has been a long history of research on active learning in machine learning community [Chapelle, 2005; Freund *et al.*, 1997]. Traditional active learning research usually considers obtaining labels to maximize some measure of predictive power or model accuracy. The most widely used measures include uncertainty sampling [Settles, 2009; Tong and Koller, 2002], estimated error reduction [Roy and McCallum, 2001] and variance reduction [Cai and He, 2012]. Recently, some researchers consider the active learning process from a perspective of data reconstruction. Transductive Experimental Design (TED) [Yu *et al.*, 2006] selects the points that the original data space can be reconstructed in a global way where each data point is linearly reconstructed by using all of the selected data points. However, given a target point, it is more reasonable to reconstruct it by using only its nearest neighbors since the points far away from the target point have little or even negative effect for the reconstruction.

In this paper we propose a novel method, called *Active Learning via Neighborhood Reconstruction* (ALNR) to select the most informative points by exploring the local geometrical structure of data space directly. Specifically, each data point can be reconstructed by only using the selected points in its neighborhood. Two important regularization penalties are considered in the objective function to incorporate the locality information and enforce the sparsity of the coefficients for final reconstruction separately. Furthermore, we propose an efficient two-stage iterative scheme to solve the final optimization problem. Firstly, the optimization problem can be factored into several subproblems based on blockwise coordinate descent method. And then, by pre-defining a special tree-structure, we demonstrate that each subproblem equals to a structured sparsity-inducing regularized problem, which can be solved via a primal-dual approach efficiently. Experimental results on both synthetic and real world data sets show that our proposed ALNR indeed has better performance than other state-of-the-art active learning approaches.

The rest of this paper is organized as follows. In Sections 2, we provide a brief review of the related work about the active learning method and structured sparsity-inducing regularization. Our algorithm is introduced in section 3 and we describe the two-step iterative optimization scheme to solve the final optimization problem in the section 4. A variety of experimental results are presented in Section 5. Finally, we provide some concluding remarks in Section 6.

Notations: Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ be the set of data points, where each $\mathbf{x}_i \in \mathbb{R}^d$ corresponding to a data point. Unless specifically mentioned, we use X to represent both the matrix and set $\{\mathbf{x}_i\}$. And let $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_m] \subset \mathbf{X}$ be the set of m selected points. For any vector $\mathbf{a} = (a_1, a_2, \ldots, a_d)^T \in \mathbb{R}^d$, the l_2 norm is defined as $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{n} |a_i^2|}$, and the sup-norm of $\mathbf{a}$ is defined as $\|\mathbf{a}\|_\infty = \max_{1 \le i \le d} |a_i|$.

2 Background

The work most related to our approach is the the Transductive Experimental Design (TED) [Yu *et al.*, 2006], whose key idea is to minimize the average predictive variance of the estimated regularized linear regression function. In a geometrical view, this is equivalent to find m representative data samples $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_m] \subset \mathbf{X}$ that span a linear space to retain most of the information of $\mathbf{X}$, which can be formulated as follows

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

where α is the regularization parameter controlling the amount of shrinkage. To solve this problem, a suboptimal sequential greedy algorithm to select the m representative points one by one was proposed [Yu *et al.*, 2006] and a non-greedy algorithm is also designed for the convex relaxation of problem (1). Cai *et al.* further proposed to choose the samples in the data manifold adaptive kernel space based on the convex TED [Cai and He, 2012]. Their experimental results showed that the incorporation of the locality information can improves the performance of active learning process. Other active learning works related to our work include Simple Margin method [Tong and Koller, 2002] and LLR$_{Active}$ [Zhang *et al.*, 2011].

In sparse coding literature, the structured sparsity-inducing regularization has been introduced to enforce the sparsity in the feature vector with the consideration of the structure of the features [Roth and Fischer, 2008; Yuan and Lin, 2006]. This topic recently has attracted many researchers' attention. By assuming disjoint groups structure of features, the group lasso model was proposed to enforce sparsity on the pre-defined groups of features [Bach, 2008]. Furthermore, this model has been extended to allow groups that hierarchical as well as overlapping [Zhao *et al.*, 2009; Kim and Xing, 2010; Liu and Ye, 2010]. Considering the possible non-smoothness of the structured regularization, a series of optimization methods are also proposed to solve such problems efficiently [Jenatton *et al.*, 2011; Qin and Goldfarb, 2012].

3 The Objective Function

From formulation (1), we can see that TED reconstructs each point via a linear combination of all the selected points. Geometrically speaking, it is more reasonable to approximate $\mathbf{x}_i$ by the linear combination of only its neighbors to capture the local geometrical information of data. The recent theoretical works in machine learning [Yu *et al.*, 2009; Wang *et al.*, 2010] have shown that the learning performance can be significantly enhanced if the local geometrical structure is exploited. For any selected point $\mathbf{z}_j \in \mathbf{Z}$, we denote function $d(\mathbf{z}_j, \mathbf{x}_i)$ to be the distance between $\mathbf{z}_j$ and $\mathbf{x}_i$, where $d(\cdot,\cdot)$ can be any distance such as Euclidean distance and geodesic distance. Intuitively, the smaller $d(\mathbf{z}_j, \mathbf{x}_i)$ is ($\mathbf{z}_j$ is closer to $\mathbf{x}_i$), the greater effect $\mathbf{z}_j$ should have for the local reconstruction of $\mathbf{x}_i$ and vice versa.

Motivated by these facts, we propose a new novel method called *Active Learning via Neighborhood Reconstruction* (ALNR) to select the most informative points. For each point $\mathbf{x}_i$, we claim that the corresponding reconstruction should be built mainly over its neighborhood. By penalizing the coefficients of the reconstruction, we formulate our objective function as follows

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{A}} \quad & \sum_{i=1}^{n}(\|\mathbf{x}_i - \mathbf{Z}\mathbf{a}_i\|_2^2 + \mu\sum_{j=1}^{m}|a_{ji}|d(\mathbf{z}_j, \mathbf{x}_i)) \\ s.t. \quad & \mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_m] \subset \mathbf{X}, \\ & \mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n] \in \mathbb{R}^{m\times n}, \end{aligned} \tag{2}$$

where a_{ji} is the j-th element of vector $\mathbf{a}_i$ and μ is a regularization parameter. In the objective function (2), the first term $\|\mathbf{x}_i - \mathbf{Z}\mathbf{a}_i\|_2^2$ means that $\mathbf{x}_i$ should be close to its physical approximation $\mathbf{Z}\mathbf{a}_i$, and the second term $\sum_{j=1}^{m}|a_{ji}|d(\mathbf{z}_j, \mathbf{x}_i)$ restricts the reconstruction of $\mathbf{x}_i$ to be localized.

Unfortunately, the optimization problem (2) is a combinatorial problem. The optimal representative data set for $\mathbf{x}_i$ is usually not optimal for other points. To get the reconstructions of all the data points, we would have to search over an exponential number of possible sets to determine the unique optimal $\mathbf{Z}$.

Considering this difficulty, we firstly relax the problem (2) by assuming that all the data points are selected to be representative points, i.e., $\mathbf{Z} = \mathbf{X}$. In this case, we transform problem (2) to a special case of sparse coding problem [Xie *et al.*, 2010], which can be formulated as follows

$$\begin{aligned} \min_{\mathbf{A}} \quad & \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \mu\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ji}|d(\mathbf{x}_j, \mathbf{x}_i) \\ s.t. \quad & \mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_n] \in \mathbb{R}^{n\times n}, \end{aligned} \tag{3}$$

where $\|\cdot\|_F$ stands for the Frobenius norm of matrices. Since our target is to choose the m most informative points, the corresponding coefficients of linear reconstructions on these m selected points must have larger weights and the weights of each data point on other $n - m$ points should be as small as possible. Notice that the l-th row of matrix $\mathbf{A}$ reflects the importance of the point $\mathbf{x}_l$ in the linear reconstruction of original data space. This is equivalent to require the optimal solution $\mathbf{A}$ should be sparse enough in rows. And the m most informative rows of $\mathbf{A}$ are exactly corresponding to the finally selected m representative data points of the whole data set $\mathbf{X}$.

To enforce the sparsity of the row vectors of the final optimal solution $\mathbf{A}$, we propose to utilize the sup-norm $\|\cdot\|_\infty$ to penalize each row vector. Sup-norm has the effect of "grouping" the elements in vector such that they can achieve zeros simultaneously. For simplicity, we define weight matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$, where $\mathbf{D}_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$. We further denote vector $\tilde{\mathbf{a}}_i \in \mathbb{R}^n$ where $\tilde{\mathbf{a}}_i^T$ is the i-th row vector of matrix $\mathbf{A}$ and $\tilde{a}_{ij}$ to be the j-th element of vector $\tilde{\mathbf{a}}_i$, i.e. $\tilde{a}_{ij} = a_{ji}$. Then we can reformulate our final objective function as follows

$$\begin{aligned} \min_{\mathbf{A}} \quad & \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \mu\sum_{i=1}^{n}\sum_{j=1}^{n}|\tilde{a}_{ij}|\mathbf{D}_{ij} + \lambda\sum_{i=1}^{n}\|\tilde{\mathbf{a}}_i\|_\infty \\ s.t. \quad & \mathbf{A}^T = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, ..., \tilde{\mathbf{a}}_n] \in \mathbb{R}^{n\times n}, \end{aligned} \tag{4}$$

where μ and λ are two positive trade-off parameters to control the degree of penalty.

Once the optimal solution $[\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \ldots, \tilde{\mathbf{a}}_n]$ of the original optimization problem (4) is obtained, we rank all the data points according to the value of $\|\tilde{\mathbf{a}}_s\|_\infty (s = 1, 2, \ldots, n)$ in descending order and the top m points are selected.

4 Optimization Method

In this section, we discuss how to solve the optimization problem (4). Although the two regularizations in the objective function are both convex individually, the main challenge is how to deal with them simultaneously. Notice that the objective function is separable, we propose a two-stage iterative optimization scheme to solve problem (4), where the original problem is factored into n subproblems based on blockwise coordinate descent method in the first step and each subproblem can be solved by using structured optimization techniques efficiently in the last step.

4.1 Blockwise Coordinate Descent

Recall that the $\tilde{\mathbf{a}}_i$ in problem (4) represents the coefficient vector of the i-th sample reconstructing all the n samples, we call $\tilde{\mathbf{a}}_i$ a block. Since the objective function is separable, the blockwise coordinate descent method consists of simultaneously updating the coefficients within each block while holding all the others fixed, then cycling through this process. Therefore, if the current estimates are $\tilde{\mathbf{a}}_i, i = 1, ..., n$, then $\tilde{\mathbf{a}}_i$ is updated by the following subproblems:

$$\tilde{\mathbf{a}}_i^{new} \longleftarrow \arg\min_{\tilde{\mathbf{a}}_i} \left(F(\tilde{\mathbf{a}}_i) = f(\tilde{\mathbf{a}}_i) + \Phi(\tilde{\mathbf{a}}_i) \right), \quad (5)$$

where the first term $f(\tilde{\mathbf{a}}_i) = \|\mathbf{R}_i - \mathbf{x}_i \tilde{\mathbf{a}}_i^T\|_F^2$ and $\mathbf{R}_i = \mathbf{X} - \sum_{j \neq i} \mathbf{x}_j \tilde{\mathbf{a}}_j^T$ denotes the partial residual matrix. And the second term $\Phi(\tilde{\mathbf{a}}_i) = \sum_{j=1}^n \mu |\tilde{a}_{ij}| \mathbf{D}_{ij} + \lambda \|\tilde{\mathbf{a}}_i\|_\infty$ is the penalty for each subproblem.

If the trade off parameter $\mu = 0$, the penalty term $\Phi(\tilde{\mathbf{a}}_i)$ turns into the general sup-norm. Then the problem in (5) decouples into a sup-norm penalized least squares regression problem. It has been shown that there exists a closed form solution for this type of problem [Liu *et al.*, 2009]. However, when $\mu \neq 0$, the penalty term $\Phi(\tilde{\mathbf{a}}_i)$ has a more complex structure, which leads a much more sophisticated situation than before. So the most critical part in our optimization is how to deal with $\Phi(\tilde{\mathbf{a}}_i)$ in subproblem (5) efficiently.

4.2 Geometrical Interpretation of Φ

In this subsection, we show that the penalty term $\Phi(\tilde{\mathbf{a}}_i)$ actually can be described as a hierarchical sparsity-inducing regularization with a predefined tree-structured set of groups, which can be defined as follows

Definition 4.1. ***(Tree-structured set of groups** [Jenatton* et al.*, 2011]) A set of groups $\mathcal{G} = \{g\}_{g \in \mathcal{G}}$ is said to be tree-structured in $\{1, \ldots, n\}$, if $\bigcup_{g \in \mathcal{G}} g = \{1, \ldots, n\}$ and for all $g, h \in \mathcal{G}$,*

$$(g \bigcap h \neq \emptyset) \Longrightarrow (g \subseteq h \textit{ or } h \subseteq g).$$

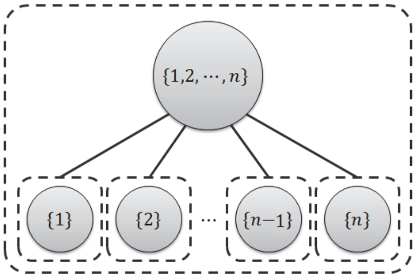

Figure 1: The illustration of our two-layer tree-structured set of groups $\mathcal{M} = \{\{1, 2, \ldots, n\}, \{1\}, \{2\}, \ldots, \{n\}\}$. The root node is assigned to be the group $\{1, 2, \ldots, n\}$, and the j-th leaf node is assigned to be the group $\{j\}$ separately, $j = 1, 2, \ldots, n$.

For such a set of groups, there exists a (non-unique) total order relation $\preceq$ such that

$$(g \preceq h) \Longrightarrow (g \subseteq h \textit{ or } g \bigcap h = \emptyset).$$

Based on this definition, given a tree-structured set of groups $\mathcal{G} = \{g\}_{g \in \mathcal{G}}$, for any vector $\mathbf{b} \in \mathbb{R}^n$, the hierarchical sparsity-inducing regularization is defined as follows

$$\Omega(\mathbf{b}) \triangleq \sum_{g \in \mathcal{G}} \omega_g \|\mathbf{b}_{|g}\|, \quad (6)$$

where $\mathbf{b}_{|g} \in \mathbb{R}^n$ whose coordinates are equal to those of $\mathbf{b}$ for indices in the group g, and 0 otherwise. Specifically, $\|\cdot\|$ stands for the l_∞ or l_2 norm, and $(\omega_g)_{g \in \mathcal{G}}$ denotes some predefined weights. By the theoretical analysis of [Zhao *et al.*, 2009], when penalizing by Ω, some of the vectors $\mathbf{b}_{|g}$ are set to zero for some $g \in \mathcal{G}$, which leads the desired effect of structured sparsity.

With the notations of tree-structured set $\mathcal{G}$ and its associated hierarchical sparsity-inducing regularization, we construct a two-layer tree-structure of groups (see Figure 1) for the specific configuration of $\Phi(\tilde{\mathbf{a}}_i)$ in subproblem (5) . The root node in the first layer is assigned with the group $g_{n+1} = \{1, 2, \ldots, n\}$, and the n leaf nodes in the second layer are assigned with the group $g_j = \{j\}$ separately, $j = 1, \ldots, n$. It is easy to check that the set $\mathcal{M} = \{g_1, g_2, \ldots, g_{n+1}\}$ is a tree-structured set of groups according to the Definition 4.1. Then the associated hierarchical sparsity-inducing regularization of $\mathcal{M}$ can be formulated as

$$\sum_{j=1}^{n+1} \omega_j \|\tilde{\mathbf{a}}_{i|g_j}\|_\infty, \quad (7)$$

where the weights are set to be

$$\omega_{n+1} = \lambda, \quad \omega_j = \mu D_{ij}, \quad j = 1, \ldots, n. \quad (8)$$

According to the definition of $\tilde{\mathbf{a}}_{i|g_j}$, it is obvious to see that

$$\|\tilde{\mathbf{a}}_{i|g_{n+1}}\|_\infty = \|\tilde{\mathbf{a}}_i\|_\infty, \quad \|\tilde{\mathbf{a}}_{i|g_j}\|_\infty = |\tilde{a}_{ij}|, \quad j = 1, \ldots, n.$$

Based on above results, we can find that

$$\Phi(\tilde{\mathbf{a}}_i) = \sum_{j=1}^n \mu |\tilde{a}_{ij}| \mathbf{D}_{ij} + \lambda \|\tilde{\mathbf{a}}_i\|_\infty = \sum_{j=1}^{n+1} \omega_j \|\tilde{\mathbf{a}}_{i|g_j}\|_\infty. \quad (9)$$

Algorithm 1 Active Learning based on Neighborhood Reconstruction

```
Input:
    • The candidate data set: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$
    • The number of selected points: $m$
    • The parameters: $\mu$, $\lambda$ and $N$
Output:
    • The set of $m$ selected points $\mathbf{Z}$
Compute the weight matrix $D$.
for $k = 1, \ldots, N$ do
    for $i = 1, \ldots, n$ do
        Compute the weights of $\Phi(\tilde{\mathbf{a}}_i)$ according to (8).
        Update $\tilde{\mathbf{a}}_i^{k+1}$ as in (15) until convergence.
    end for
end for
Rank the data points according to $\|\tilde{\mathbf{a}}_s\|_\infty$ $(s = 1, \ldots, n)$
in descending order, and return the top $m$ data points.
```

So penalty term $\Phi(\tilde{\mathbf{a}}_i)$ is just the hierarchical sparsity-inducing regularization of the tree-structured set $\mathcal{M}$.

4.3 Proximal Method for Subproblem (5)

Recall the fact that in the objective function $F(\tilde{\mathbf{a}}_i)$ of problem (5), $f(\tilde{\mathbf{a}}_i)$ is convex and differentiable whereas the penalty term $\Phi(\tilde{\mathbf{a}}_i)$ is convex and nondifferentiable with respect to $\tilde{\mathbf{a}}_i$. We propose to solve the problem (5) by using proximal method [Nesterov, 2007; Beck and Teboulle, 2009], which has been widely applied for its outstanding ability to deal with large-scale, possibly nonsmooth problems.

Proximal method solves the problem (5) iteratively. Firstly, for any $t > 0$, in the k-th iteration, the proximal method constructs an approximation of $F(\tilde{\mathbf{a}}_i)$ at the current estimate $\tilde{\mathbf{a}}_i^k$ as follows

$$\begin{aligned} Q(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_i^k) = f(\tilde{\mathbf{a}}_i^k) + \langle \tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_i^k, \nabla f(\tilde{\mathbf{a}}_i^k) \rangle + \frac{1}{2t} \|\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_i^k\|_2^2 \\ + \Phi(\tilde{\mathbf{a}}_i). \end{aligned} \tag{10}$$

Then, we update $\tilde{\mathbf{a}}_i^{k+1}$ as the unique minimizer of $Q(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_i^k)$:

$$\begin{aligned} \tilde{\mathbf{a}}_i^{k+1} &= \arg\min_{\tilde{\mathbf{a}}_i \in \mathbb{R}^n} Q(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_i^k) \\ &= \arg\min_{\tilde{\mathbf{a}}_i \in \mathbb{R}^n} \frac{1}{2t} ||\tilde{\mathbf{a}}_i - (\tilde{\mathbf{a}}_i^k - t\nabla f(\tilde{\mathbf{a}}_i^k))||_2^2 + t\Phi(\tilde{\mathbf{a}}_i). \end{aligned} \tag{11}$$

Notice that $\Phi(\tilde{\mathbf{a}}_i)$ is the hierarchical sparsity-inducing regularization associated with tree-structured set of groups $\mathcal{M}$ and the dual norm of sup-norm is l_1 norm, the problem (11) can be solved via a primal-dual approach based on the theoretical work of Jenatton *et al.* Specifically, the detailed formulation of the dual problem of (11) can be described in the following theorem.

Theorem 4.1. *[Jenatton* et al.*, 2011] Let us consider the following problem*

$$\begin{aligned} &\min_{\xi} \ \|(\tilde{\mathbf{a}}_i^k - t\nabla f(\tilde{\mathbf{a}}_i^k)) - \sum_{j=1}^{n+1} \xi_{g_j}\|_2^2 - \|\tilde{\mathbf{a}}_i^k - t\nabla f(\tilde{\mathbf{a}}_i^k)\|_2^2 \\ &s.t.\ \forall j \in \{1, 2, \ldots, n+1\},\ \|\xi_{g_j}\|_1 \leq t\omega_j \ and \\ &\qquad \xi_{g_j,l} = 0 \ if\ l \notin g_j. \end{aligned} \tag{12}$$

where $\xi = [\xi_{g_1}, \xi_{g_2}, \ldots, \xi_{g_{n+1}}] \in \mathbb{R}^{n \times (n+1)}$ *and* $\xi_{g_j,l}$ *denotes the l-th coordinate of the vector* $\xi_{g_j} \in \mathbb{R}^n$. *Then the problem (11) and (12) are dual to each other and strong duality holds. In addition, the pair of primal-dual variables* $\{\tilde{\mathbf{a}}_i^*, \xi^*\}$ *is optimal if and only if* ξ^* *is a feasible point of the optimization problem (12), and*

$$\tilde{\mathbf{a}}_i^* = \tilde{\mathbf{a}}_i^k - t\nabla f(\tilde{\mathbf{a}}_i^k) - \sum_{j=1}^{n+1} \xi_{g_j}^*, \tag{13}$$

and for $\forall g_j \in \mathcal{M}$

$$\xi_{g_j}^* = \Pi_{t\omega_j}(\tilde{\mathbf{a}}_{i|g_j}^*) \ \ or \ \ \tilde{\mathbf{a}}_{i|g_j}^* = 0, \tag{14}$$

where $\Pi_{t\omega_j}(\cdot)$ *stands for the orthogonal projection onto the ball of the* l_1 *norm with radius* $t\omega_j$.

Based on this theorem, the blockwise coordinate ascent method is used to solve the dual problem (12) efficiently. For each $g_j \in \mathcal{M}$, we update the vector ξ_{g_j} and keep other dual variables fixed. Then the primal variable $\tilde{\mathbf{a}}_i^{k+1}$ and dual variable ξ_{g_j} are updated alternatively as follows

$$\begin{cases} \tilde{\mathbf{a}}_i^{k+1} \longleftarrow \tilde{\mathbf{a}}_i^k - t\nabla f(\tilde{\mathbf{a}}_i^k) - \sum\limits_{l=1, l \neq j}^{n+1} \xi_{g_l}, \\ \xi_{g_j} \longleftarrow \Pi_{t\omega_j}(\tilde{\mathbf{a}}_{i|g_j}^{k+1}). \end{cases} \tag{15}$$

This process is repeated until a stable $\tilde{\mathbf{a}}_i^{k+1}$ is obtained. Based on the recent theoretical work of structured optimization method, the optimal $\tilde{\mathbf{a}}_i^{k+1}$ can be obtained exactly with only one iterations over all the groups of $\mathcal{M}$ [Jenatton *et al.*, 2011].

Overall, we summarize the complete procedure in Algorithm 1, whose convergence can be guaranteed by the blockwise coordinate descent method.

5 Experiments

To demonstrate the effectiveness of our proposed algorithm, we evaluate and compare four active learning methods:

- **Random Sampling** method which randomly selects points from the data set. This method is used as the baseline for active learning.
- **Simple Margin** method which selects the points closest to the current decision boundary of the SVM classifier as the most informative ones [Tong and Koller, 2002].
- **Transductive Experimental Design** (TED) is proposed in [Yu *et al.*, 2006].

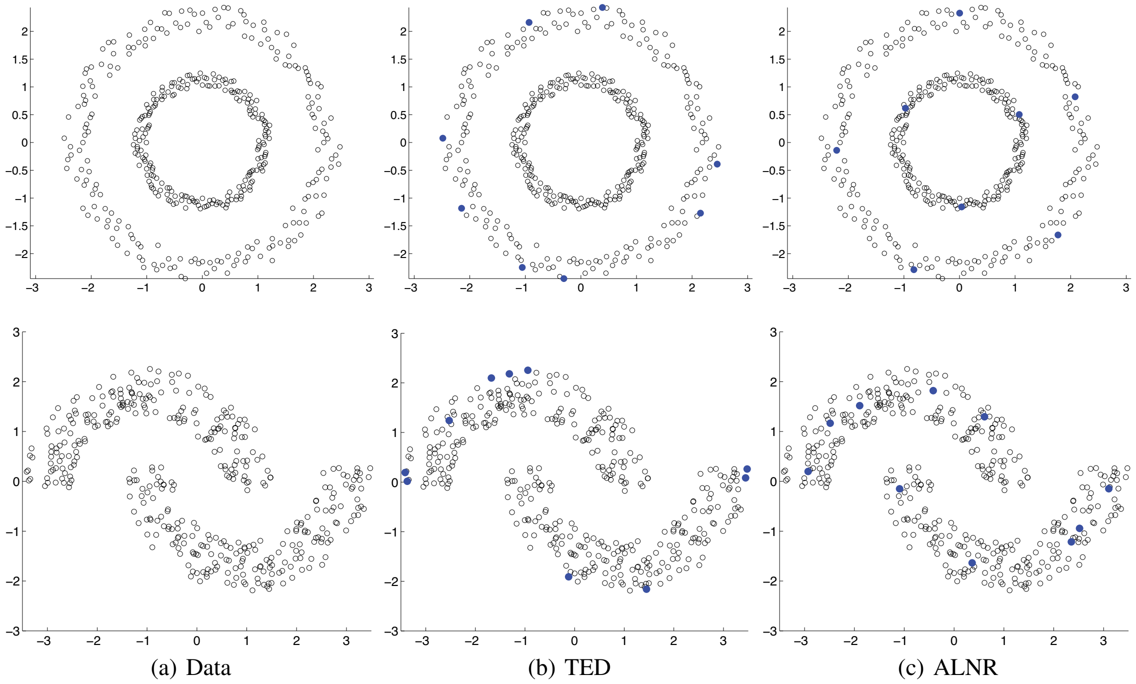

Figure 2: Data selection by active learning algorithms TED and ALNR on two-circle data set and two-moon data set. The selected data points are marked as solid dots. Clearly, on both data sets, the points selected by our proposed ALNR algorithm can better represent the original data set.

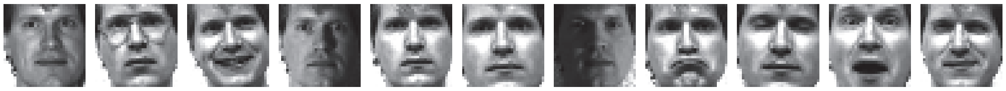

Figure 3: The sample cropped face images of one individual from Yale database. The variations contains different lighting condition, face expression, and with/without glasses.

- **Active Learning based on Neighborhood Reconstruction** (ALNR) proposed in this paper. The data is first clustered by a simple spectral clustering method. Then for any two data points x, y, the *dissimilarity* measure $d(x, y)$ is set to be the geodesic distance between x and y when x and y are in a same cluster, and $d(x, y)$ is set to be ∞ when x and y are in different clusters.

We note that all the methods use linear SVM with squared hinge loss function as the base classification method.

5.1 Toy examples

In this subsection, we apply the active learning algorithms on two synthetic data sets to give an intuitive idea of how ALNR performs differently from TED. The synthetic data sets are described as follows

- Two-circle data set (Figure 2): There are two circles, each contains 200 points.
- Two-moon data set (Figure 2): There are two moons, each contains 200 points.

Then we apply TED and ALNR to select the most informative points on the two synthetic data sets. The results are shown in Figure 2. The points selected by each active learning algorithm are marked as solid blue dots. Compared with TED, the points selected by ALNR can better represent the original data set. This is because ALNR can better capture the nonlinear structure of the data by incorporating the important local geometrical information into the objective function (4).

5.2 Face Recognition

In this subsection, we investigate the performance of the different active learning algorithms by using them to solve the face recognition problem on Yale face database.

Yale face database contains 165 gray scale images of 15 individuals. There are 11 images per subject, including variations in lighting condition (left-light, center-light, right-light), facial expression (normal, happy, sad, sleepy, surprised, and wink), and with/without glasses. All the images are manually aligned and cropped. The size of each cropped image is 32×32 pixels, with 256 gray levels per pixel. Thus each image is represented as a 1024-dimensional vector. So face recognition is a classification problem in a 1024-dimensional Euclidean Space. Figure 3 shows some sample images from the Yale face database.

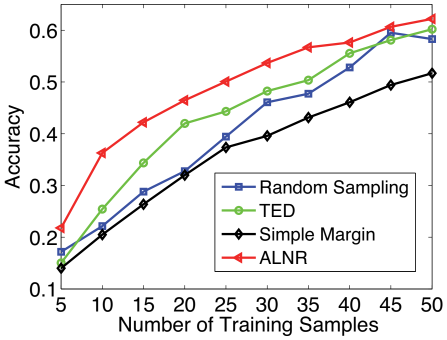

Figure 4: The average classification accuracy versus the number of training samples.

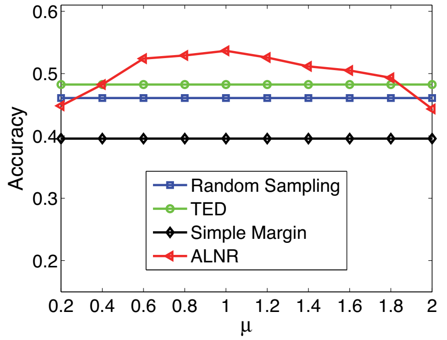

Figure 5: The performance of ALNR versus the parameter μ. When λ is set to 0.5, it can be shown that ALNR achieves consistent good performance with the μ varying from 0.4 to 1.8.

The evaluations are conducted with 10 randomly generated subsets of the original data set. The average classification accuracy is computed over these 10 tests. For each test, 10 images from each class are randomly chosen to form the data set. Therefore, there are 150 (15×10) face images per test, and each active learning algorithm is applied to select a given number $\omega = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ of training samples. The unselected samples are used as the testing data.

Figure 4 shows the average classification accuracy versus the number of training (selected) samples of different active learning algorithm on Yale face database. It can be seen that, for all compared algorithms, the classification accuracy increases with the size of training examples. And our proposed ALNR has the best performance over all the size of train examples. It is worthwhile to note that ALNR performs especially good when the number of training examples is limited which is very common in the practical applications. Our experimental results also demonstrate that the local geometrical information can greatly improve the performance of active learning process.

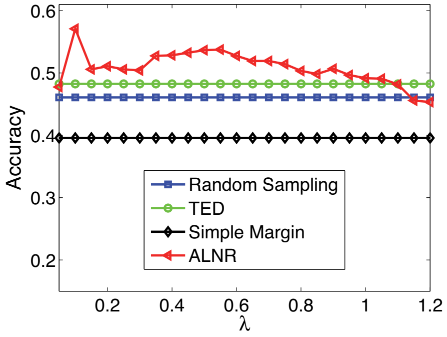

Figure 6: The performance of ALNR versus the parameter λ. When μ is set to 1, it can be shown that ALNR always has a good performance with λ varying from 0 to 1.

5.3 Parameter Selection

There are two essential parameters, μ and λ, in our ALNR algorithm, where parameter μ is used to control the locality and λ is used to control the degree of sparsity. In our previous experiments, we simply set $\lambda = 0.5, \mu = 1$. In this subsection, we examine how the average performance of ALNR varies with the parameters μ and λ separately. We conduct 10 random tests as in the last subsection and the number of selected training examples ω is set to be 30. When λ is fixed to be 0.5, the impact of μ for average performance is shown in Figure 5, where we can see ALNR can achieve consistent good performance with μ varying from 0.4 to 1.8. And Figure 6 shows the experimental results with μ fixed to be 1 and λ varying from 0 to 1, where ALNR always has a good performance with different λ.

6 Conclusion

In this paper, we propose a novel method called *Active Learning via Neighborhood Reconstruction* (ALNR) to select the most representative points from a local reconstruction perspective. The reconstruction of each data point is mainly conducted over the selected points only in its neighborhood. In this way, we incorporate the important local geometrical information into the active learning process. An efficient two-stage iterative scheme is also proposed for the final optimization problem. Experimental results on two synthetic and one real world data sets show the effectiveness of our approach. It is interesting to explore how to accelerate ALNR for real world applications in our following work.

Acknowledgments

This work is supported by National Basic Research Program of China (973 Program) under Grant 2012CB316400 and National Natural Science Foundation of China (Grant No: 61125203, 61222207, 61233011, 90920303).

References

[Bach, 2008] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

[Chapelle, 2005] Olivier Chapelle. Active learning for parzen window classifier. *In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[Freund *et al.*, 1997] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[Gosselin and Cord, 2008] P. H. Gosselin and M. Cord. Active learning methods for interactive image retrieval. *IEEE Transactions on Image Processing*, 17(7):1200–1211, 2008.

[He *et al.*, 2012] Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. Document summarization based on data reconstruction. In *the 26th AAAI Conference on Artificial Intelligence*, 2012.

[Jenatton *et al.*, 2011] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 2011.

[Kapoor *et al.*, 2010] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision*, 88(2):169–188, June 2010.

[Kim and Xing, 2010] Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine learning*, 2010.

[Liu and Ye, 2010] Jun Liu and Jieping Ye. Moreau-yosida regularization for grouped tree structure learning. In *Advances in Neural Information Processing Systems*, 2010.

[Liu *et al.*, 2009] Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[Nesterov, 2007] Yurii Nesterov. Gradient methods for minimizing composite objective function. Technical report, 2007.

[Qi *et al.*, 2008] Guojun Qi, Xiansheng Hua, Yong Rui, Jinhui Tang, and Hongjiang Zhang. Two-dimensional active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[Qin and Goldfarb, 2012] Zhiwei Qin and Donald Goldfarb. Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, 2012.

[Roth and Fischer, 2008] Volker Roth and Bernd Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine learning*, 2008.

[Roy and McCallum, 2001] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

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

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

[Wang *et al.*, 2010] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[Xie *et al.*, 2010] Bo Xie, Mingli Song, and Dacheng Tao. Large-scale dictionary learning for local coordinate coding. In *British Machine Vision Conference,Aberystwyth, UK*, 2010.

[Yan *et al.*, 2003] Rong Yan, Jie Yang, and Alexander Hauptmann. Automatically labeling video data using multi-class active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 516–523, 2003.

[Yu *et al.*, 2006] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[Yu *et al.*, 2009] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *Advances in Neural Information Processing Systems*, pages 2223–2231, 2009.

[Yuan and Lin, 2006] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[Zhang *et al.*, 2011] Lijun Zhang, Chun Chen, Jiajun Bu, Deng Cai, Xiaofei He, and Thomas S. Huang. Active learning based on locally linear reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2026–2038, 2011.

[Zhao *et al.*, 2009] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.