Neural Comput & Applic (2013) 23:2085–2092
DOI 10.1007/s00521-012-1155-3

ORIGINAL ARTICLE

Neighborhood preserving D-optimal design for active learning and its application to terrain classification

Yingjie Gu · Zhong Jin

Received: 21 December 2011 / Accepted: 28 August 2012 / Published online: 23 September 2012

Abstract In many real-world applications, labeled data are usually expensive to get, while there may be a large amount of unlabeled data. To reduce the labeling cost, active learning attempts to discover the most informative data points for labeling. The challenge is which unlabeled samples should be labeled to improve the classifier the most. Classical optimal experimental design algorithms are based on least-square errors over the labeled samples only while the unlabeled points are ignored. In this paper, we propose a novel active learning algorithm called neighborhood preserving D-optimal design. Our algorithm is based on a neighborhood preserving regression model which simultaneously minimizes the least-square error on the measured samples and preserves the neighborhood structure of the data space. It selects the most informative samples which minimize the variance of the regression parameter. We also extend our algorithm to nonlinear case by using kernel trick. Experimental results on terrain classification show the effectiveness of proposed approach.

Keywords Active learning · Terrain classification · Optimal experimental design (OED) · Neighborhood preserving

1 Introduction

For large-scale real problems, such as image retrieval and terrain classification, there are large numbers of unlabeled data, while the labeled data are usually difficult to get. Semi-supervised learning [1–3] makes full use of the information from both labeled and unlabeled data to address this problem. Besides, active learning [4, 5] queries some instances for manual labeling to construct a training set. In recent years, active learning has gained increasing interests and demonstrated its effectiveness in various applications.

In statistics, the problem of selecting samples to label is usually referred to as experimental design. The sample $\mathbf{x}$ is referred to as experiment and its label y is referred to as measurement. Optimum experimental design (OED) [6–8] tries to minimize the variance of a parameter model. Traditional experimental design approaches include A-optimal design, D-optimal design, and E-optimal design. But none of them explore additional information contained in the unlabeled data.

Besides OED-based active learning approaches, another important method of active learning is based on support vector machines ($\text{SVM}_{\text{active}}$) [9, 10]. $\text{SVM}_{\text{active}}$ asks the user to label the points which are closet to the decision boundary. The disadvantage of $\text{SVM}_{\text{active}}$ is that the estimated boundary may not be accurate enough especially when the number of training examples is small. Moreover, since it needs an initial decision boundary, it cannot be applied when there are no labeled data points. Some other SVM-based active learning algorithms can be found in [11, 12].

Recently, He [13], Chen [14], and Zhang [15] have proposed new active learning methods called LapRDD, LapGOD, and CLapRID, respectively. All of them are based on Laplacian regularized least squares (LapRLS) [3, 16] and using different optimality criteria of experimental design. For example, He used graph Laplacian in the formulation of D-optimal design while Chen used

Y. Gu (✉) · Z. Jin
School of Computer Science and Engineering,
Nanjing University of Science and Technology,
Nanjing, Jiangsu 86, China
e-mail: csyjgu@gmail.com

graph Laplacian in the formulation of G-optimal design. Besides, Zhang [17] proposed a new active learning method based on locally linear reconstruction which outperforms the classical algorithms. Shen [18] introduced the idea of column subset selection, which aims to select the most representation columns from a data matrix, into active learning and propose a novel active learning algorithm called $\text{CSS}_{\text{active}}$.

Motivated by recent progresses of neighborhood preserving regression [19] and optimal experimental design, we propose a novel active learning algorithm called neighborhood preserving D-optimal design (NPDOD). Unlike traditional experimental design methods whose loss functions are only defined on the measured points, the loss function of our proposed algorithm is defined on both measured and unmeasured points. Specifically, we use a loss function which imposes locally linear reconstruction error into the standard least-square-error-based loss function. Using techniques from experimental design, we can select the most informative data points which are presented to the user for labeling.

The rest of the paper is organized as follows: In Sect. 2, we provide a brief review of the related work. Our proposed NPDOD algorithm is introduced in Sect. 3. In Sect. 4, we present the nonlinear extension of our algorithm. The experimental results on terrain classification are presented in Sect. 5. Finally, we provide some suggestions for future work in Sect. 6.

2 Related work

Our algorithm is fundamentally based on neighborhood preserving regression [19]. Also, for regression-based active learning, the most related work is optimal experimental design [6], including A-optimal design, D-optimal design, and E-optimal design. In this section, we give a brief description of these approaches.

The generic problem of active learning can be formalized as follows. Given a set of data points $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, where each x_i is a d-dimensional vector, active learning aims to find a subset $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_k\} \subseteq \chi$ which contains the most informative points, that is, the subset $\mathcal{Z}$ can improve the classifier most if it is labeled and used as training data.

Throughout this paper, we use $\mathbf{x}$ to denote any point while $\mathbf{z}$ to denote the labeled point.

2.1 Optimum experimental design

We consider a linear regression model

$$y = \mathbf{w}^T\mathbf{x} + \varepsilon \tag{1}$$

where $\mathbf{w}$ is the weight vector, y is the observation, and ε is the measurement noise with zero mean and constant variance σ^2. Optimum experimental design attempts to select the most informative data points to learn a prediction function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ so that the expected prediction error can be minimized. Suppose we have a set of labeled sample points $(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_k, y_k)$ where y_i is the label of $\mathbf{z}_i$. The most popular estimation method is least squares, in which we minimize the residual sum of squares (RSS):

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^{k} \left(\mathbf{w}^T\mathbf{z}_i - y_i\right)^2 \tag{2}$$

with some simple algebraic steps, we have

$$\hat{\mathbf{w}} = \left(\mathbf{Z}^\mathbf{T}\mathbf{Z}\right)^{-1}\mathbf{Z}^\mathbf{T}\mathbf{Y} \tag{3}$$

where $\mathbf{Y} = (y_1, \ldots, y_k)^T$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]^T$. The estimate $\hat{\mathbf{w}}$ gives us an estimate of the output at a novel input: $\hat{y} = f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, and its covariance can be expressed as

$$\text{Cov}(\hat{\mathbf{w}}) = \sigma^2\left(\mathbf{Z}^\mathbf{T}\mathbf{Z}\right)^{-1} \tag{4}$$

The most informative data points are thus defined as those minimize the $\text{Cov}(\hat{\mathbf{w}})$. The criteria of OED can be classified into two categories. The first category is to select points to minimize the size of the parameter covariance matrix. The three most common criteria are

- D-optimal design: minimizes the determinant of $\text{Cov}(\hat{\mathbf{w}})$
- A-optimal design: minimizes the trace of $\text{Cov}(\hat{\mathbf{w}})$
- E-optimal design: minimizes the largest eigenvalue of $\text{Cov}(\hat{\mathbf{w}})$

Some recent work on optimal experiment design can be found in [7, 20].

In this work, we adopt the similar optimality criterion to D-optimal design for selecting the most informative data points.

2.2 Neighborhood preserving regression (NPR) [19]

Recently, Lu proposed a semi-supervised learning algorithm called neighborhood preserving regression (NPR) which is based on spectral graph theory [21–23] and locally linear embedding [24, 25]. Different from the standard regression framework which makes use of only labeled points, NPR makes use both labeled and unlabeled points. Specifically, from all the functions which can correctly classify the labeled samples, NPR selects the one which best preserves the local neighbor structure. For each sample, it may be represented as a linear combination of its p nearest neighbors. A natural assumption is that the label of this

sample can be computed by the labels of its p nearest neighbors. Let $\mathbf{W}$ be a reconstruction coefficients matrix. Thus, the NPR algorithm solves the following optimization problem:

$$L(f) = \sum_{i=1}^{k}(f(z_i) - y_i)^2 + \lambda \sum_{i=1}^{m}\left(f(x_i) - \sum_{j=1}^{m} W_{ij} f(x_j)\right)^2 \tag{5}$$

The optimal reconstruction coefficients $\mathbf{W}$ can be obtained by solving the following problem:

$$\begin{aligned} \text{Min} \quad & \sum_{i=1}^{m}\left\|\mathbf{x}_i - \sum_{j=1}^{m} W_{ij}\mathbf{x}_j\right\|^2 \\ \text{s.t.} \quad & \sum_{j=1}^{m} W_{ij} = 1, \quad i = 1, \ldots, m \\ & W_{ij} = 0 \quad \text{if } \mathbf{x}_j \notin N_p(\mathbf{x}_i) \end{aligned} \tag{6}$$

Let $\mathbf{M} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$, and $\mathbf{I}$ be a $m \times m$ identity matrix. The problem can be rewritten in the following form:

$$\text{Min} \sum_{i=1}^{k}(f(\mathbf{z}_i) - y_i)^2 + \lambda \mathbf{w}^T \mathbf{X}^{\mathbf{T}} \mathbf{M} \mathbf{X} \mathbf{w} \tag{7}$$

The optimal solution is:

$$\hat{\mathbf{w}} = (\mathbf{Z}^{\mathbf{T}}\mathbf{Z} + \lambda \mathbf{X}^{\mathbf{T}}\mathbf{M}\mathbf{X})^{-1}\mathbf{Z}^{\mathbf{T}}\mathbf{Y} \tag{8}$$

For more details about this algorithm, please see [19].

3 Neighborhood preserving D-optimal design (NPDOD)

NPR is a passive learning algorithm in which the labeled and unlabeled samples are both used, and the training samples are pregiven. We propose a new active learning algorithm which shares the similar objective function as NPR but actively selects the samples for labeling. Besides, regularized least squares [3] are adopted to the algorithm which has been proved to be more efficient than ordinary least squares [7]. This is the first work which takes into account the neighborhood structure by using both labeled and unlabeled data points in OED. Since labeling resource is usually limited in some applications, the selected data points are crucial for training a good classifier.

In this section, we introduce our active learning algorithm which is fundamentally based on NPR and regularized least squares. Thus, the regression parameter can be obtained by minimizing the following function:

$$\begin{aligned} L(f) = & \sum_{i=1}^{k}(f(\mathbf{z}_i) - y_i)^2 \\ & + \frac{\lambda_1}{2}\sum_{i=1}^{m}\left(f(\mathbf{x}_i) - \sum_{j=1}^{m} W_{ij} f(\mathbf{x}_j)\right)^2 + \lambda_2 \|f^2\| \end{aligned} \tag{9}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the trade-off parameters, y_i is the label of $\mathbf{z}_i$, and W is the reconstruction coefficient which can be obtained by (6). $\|\cdot\|$ is the vector ℓ_2-norm. The second term of the right-hand side in the cost function is the total reconstruction error. Since each sample may be represented as a linear combination of its p nearest neighbors. Therefore, the label of this sample can be computed by the labels of its p nearest neighbors. The reconstruction error should be as small as possible.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^T$ and $\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathbf{T}}(\mathbf{I} - \mathbf{W})$. The solution to minimization problem (9) is given as follows:

$$\hat{\mathbf{w}} = (\mathbf{Z}^{\mathbf{T}}\mathbf{Z} + \lambda_1 \mathbf{X}^{\mathbf{T}}\mathbf{M}\mathbf{X} + \lambda_2 \mathbf{I})^{-1}\mathbf{Z}^{\mathbf{T}}\mathbf{Y} \tag{10}$$

3.1 The objective function of NPDOD

Similar to conventional optimal experimental design techniques, we first compute the parameter covariance matrix of NPDOD. We define:

$$\mathbf{H} = \mathbf{Z}^{\mathbf{T}}\mathbf{Z} + \lambda_1 \mathbf{X}^{\mathbf{T}}\mathbf{M}\mathbf{X} + \lambda_2 \mathbf{I} \tag{11}$$

and

$$\Lambda = \lambda_1 \mathbf{X}^{\mathbf{T}}\mathbf{M}\mathbf{X} + \lambda_2 \mathbf{I} \tag{12}$$

By noticing that $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$ and $\mathbf{H}$ is symmetric, the covariance matrix of $\hat{\mathbf{w}}$ has the expression

$$\begin{aligned} \text{Cov}(\hat{\mathbf{w}}) &= \text{Cov}\left(\mathbf{H}^{-1}\mathbf{Z}^{\mathbf{T}}\mathbf{Y}\right) \\ &= \mathbf{H}^{-1}\mathbf{Z}^{\mathbf{T}}\text{Cov}(\mathbf{Y})\mathbf{Z}\mathbf{H}^{-1} \\ &= \sigma^2\mathbf{H}^{-1}\mathbf{Z}^{\mathbf{T}}\mathbf{Z}\mathbf{H}^{-1} \\ &= \sigma^2\mathbf{H}^{-1}(\mathbf{H} - \Lambda)\mathbf{H}^{-1} \\ &= \sigma^2(\mathbf{H}^{-1} - \mathbf{H}^{-1}\Lambda\mathbf{H}^{-1}) \end{aligned} \tag{13}$$

In order to make the estimator $\hat{\mathbf{w}}$ as stable as possible, the size of covariance matrix $\text{Cov}(\hat{\mathbf{w}})$ has to be as small as possible. Different measures of the size of the covariance matrix lead to different optimality criteria.

In this paper, D-optimal design is applied to select the most informative samples. Since the regularization parameters λ_1 and λ_2 are usually set to be very small, we have

$$\left|\mathbf{H}^{-1} - \mathbf{H}^{-1}\Lambda\mathbf{H}^{-1}\right| \approx \left|\mathbf{H}^{-1}\right| \tag{14}$$

So the smaller $\left|H^{-1}\right|$ is, the smaller the covariance matrix is. The problem can be rewritten as follows:

$$\max_{Z=[\mathbf{z}_1,\ldots,\mathbf{z}_k]^T} |\mathbf{H}| \tag{15}$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_k$ are selected from $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.

3.2 The algorithm

In this section, we describe a sequential optimal algorithm to solve (15), which is similar to the sequential algorithm in LapRDD [13].

Suppose a set of $k(k > 0)$ samples $\mathcal{Z}_k = \{\mathbf{z}_1, \ldots, \mathbf{z}_k\} \subset \chi$ have been selected. Let $Z_k = [\mathbf{z}_1, \ldots, \mathbf{z}_k]^T$ be a $k \times d$ matrix. We define

$$\mathbf{H}_k = \mathbf{Z}_k^T \mathbf{Z}_k + \lambda_1 \mathbf{X^T M X} + \lambda_2 \mathbf{I}, \quad k \geq 1 \tag{16}$$

and

$$\mathbf{H}_0 = \lambda_1 \mathbf{X^T M X} + \lambda_2 \mathbf{I} \tag{17}$$

The $(k+1)$th sample $\mathbf{z}_{k+1}$ can be selected by solving the following problems:

$$\mathbf{z}_{k+1} = \underset{\mathbf{z} \in \chi - \mathcal{Z}_k}{\operatorname{argmax}} \left| \mathbf{H}_k + \mathbf{z}\mathbf{z}^T \right| \tag{18}$$

By using matrix determinant lemma [26], the determinant of $\mathbf{H}_k + \mathbf{z}\mathbf{z}^T$ can be written as a multiplicative updated of the determinant of $\mathbf{H}_k$.

$$\left| \mathbf{H}_k + \mathbf{z}\mathbf{z}^T \right| = |\mathbf{H}_k| \cdot \left(1 + \mathbf{z}^T \mathbf{H}_k^{-1} \mathbf{z}\right) \tag{19}$$

Since $|\mathbf{H}_k|$ is a constant while selecting the $(k+1)$th sample, (18) can be rewritten as follows:

$$\mathbf{z}_{k+1} = \underset{z \in \chi - \mathcal{Z}_k}{\operatorname{argmax}} \mathbf{z}^T \mathbf{H}_k^{-1} \mathbf{z} \tag{20}$$

The inverse of $\mathbf{H_{k+1}}$ can be updated based on the inverse of $\mathbf{H}_k$ after the $(k+1)$th sample is selected. By using the Sherman–Morrison formula

$$\begin{aligned} \mathbf{H}_{k+1}^{-1} &= (\mathbf{H}_k + \mathbf{z}_{k+1} \mathbf{z}_{k+1}^{\mathbf{T}})^{-1} \\ &= \mathbf{H}_k^{-1} - \frac{\mathbf{H}_k^{-1} \mathbf{z}_{k+1} \mathbf{z}_{k+1}^T \mathbf{H}_k^{-1}}{1 + \mathbf{z}_{k+1}^T \mathbf{H}_k^{-1} \mathbf{z}_{k+1}} \end{aligned} \tag{21}$$

Above analysis shows that we do not need to compute the determinant and inverse of covariance matrix. Instead, at each iteration, we select a new sample $\mathbf{z}$ such that maximize $\mathbf{z}^T \mathbf{H}_k^{-1} \mathbf{z}$ and $\mathbf{H}_k^{-1}$ can be updated efficiently in terms of (21). The sequential approach is summarized in Table 1.

4 Nonlinear neighborhood preserving D-optimal design

Most of traditional optimal experimental design techniques are based on linear regression model. However, in many real-world applications, the data may not be linearly separable. In this section, we discuss how to generalize our NPDOD algorithm to nonlinear case by performing experimental design in reproducing kernel Hilbert space (RKHS).

Let $\mathcal{K}(\cdot,\cdot)$be a positive definite kernel, and $\mathcal{F}$ be the corresponding reproducing kernel space. Several popular kernel functions are Gaussian kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{-\|\mathbf{x}_i - \mathbf{x}_j^2\|}{\sigma^2})$; polynomial kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = (1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^d$; Sigmoid kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + \alpha)$.

Table 1 The sequential approach for NPDOD

Input: The candidate data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, the number of nearest neighbor p, the number(k) of samples to be selected and the parameter λ_1 and λ_2

Output: The indexes of the k most representative samples, $\mathcal{Z} = \{\mathbf{z}_I, \ldots, \mathbf{z}_k\}$

1: Initialize $\mathbf{W}$ by solving problem (6)
2: $\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathbf{T}}(\mathbf{I} - \mathbf{W})$
3: $\mathcal{Z} \leftarrow \emptyset$
4: $\mathbf{H} = \lambda_1 \mathbf{X^T M X} + \lambda_2 \mathbf{I}$
5: for $n = 1$ to k do
6: for $i = 1$ to m do
7: if i $\notin \mathcal{Z}$ then
8: $A(i) = \mathbf{x}_i^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{x}_i$
9: end if
10: end for
11: $sn = \operatorname{argmax}_{i \notin \mathcal{Z}} A(i)$
12: $\mathcal{Z} \leftarrow \mathcal{Z} \bigcup sn$
13: $\mathbf{H}^{-1} \leftarrow \mathbf{H}^{-1} - \frac{\mathbf{H}^{-1} \mathbf{x}_{sn} \mathbf{x}_{sn}^T \mathbf{H}^{-1}}{1 + \mathbf{x}_{sn}^T \mathbf{H}^{-1} \mathbf{x}_{sn}}$
14: end for
15: return $\mathcal{Z}$

Then, we seek a function $f \in \mathcal{F}$ such that the following objective function is minimized:

$$\begin{aligned} L(f)_{f \in \mathcal{F}} = & \sum_{i=1}^{k} (f(\mathbf{z}_i) - y_i)^2 \\ & + \frac{\lambda_1}{2} \sum_{i=1}^{m} \left(f(\mathbf{x}_i) - \sum_{j=1}^{m} W_{ij} f(\mathbf{x}_j) \right)^2 + \lambda_2 \|f\|_{f \in \mathcal{F}}^2 \end{aligned} \tag{22}$$

The representer theorem [3] can be used to show that the solution is an expansion of kernel functions over both the labeled and unlabeled data.

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{m} \hat{\alpha}_i \varphi(\mathbf{x}_i) \tag{23}$$

Let $\hat{\alpha} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_m]^T$. The optimal solution is

$$\hat{\alpha} = (\mathbf{K_{XZ}} \mathbf{K_{ZX}} + \lambda_1 \mathbf{K_{XX}} \mathbf{M} \mathbf{K_{XX}} + \lambda_2 \mathbf{K_{XX}})^{-1} \mathbf{K_{XZ}} \mathbf{Y} \tag{24}$$

with covariance

$$\mathrm{Cov}(\hat{\alpha}) \approx \sigma^2 (\mathbf{K_{XZ}} \mathbf{K_{ZX}} + \lambda_1 \mathbf{K_{XX}} \mathbf{M} \mathbf{K_{XX}} + \lambda_2 \mathbf{K_{XX}})^{-1} \tag{25}$$

where $\mathbf{K_{XZ}}$ is a $m \times k$ matrix with $K_{XZ,ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{z}_j)$, K_{XX} is a $m \times m$ matrix with $K_{XZ,ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{K_{ZX}} = \mathbf{K_{XZ}^T}$. The nonlinear problem is defined as follows:

$$\max_{\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]^T} |\mathbf{K_{XZ}} \mathbf{K_{ZX}} + \lambda_1 \mathbf{K_{XX}} \mathbf{M} \mathbf{K_{XX}} + \lambda_2 \mathbf{K_{XX}}| \tag{26}$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_k$ are selected from $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Let $\mathbf{u}_i$ be the ith column vector of $\mathbf{K_{XX}}$, and $\mathcal{U}$ be the set of

$\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$. Clearly, $\mathbf{u}_i = (\mathcal{K}(\mathbf{x}_i, \mathbf{x}_1), \ldots, \mathcal{K}(\mathbf{x}_i, \mathbf{x}_m))^T$. Similar to selecting $\mathbf{x}_i$ in the original space, here we select $\mathbf{u}_i$ in kernel space.

Suppose a set of $k(k > 0)$ samples $\mathcal{V}_k = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\} \subset \mathcal{U}$ have been selected. Let $\mathbf{V_k} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]^T$ be a $k \times d$ matrix. We define

$$\mathbf{P_k} = \mathbf{K_{XV_k}} \mathbf{K_{V_kX}} + \lambda_1 \mathbf{K_{XX}} \mathbf{M} \mathbf{K_{XX}} + \lambda_2 \mathbf{K_{XX}} \tag{27}$$

and

$$\mathbf{P}_0 = \lambda_1 \mathbf{K_{XX}} \mathbf{M} \mathbf{K_{XX}} + \lambda_2 \mathbf{K_{XX}} \tag{28}$$

The $k + 1$th sample $\mathbf{v}_{k+1}$ can be selected by solving the following problems:

$$\mathbf{v}_{k+1} = \underset{\mathbf{v} \in \mathcal{U} - \mathcal{V}_k}{\operatorname{argmax}} \left| \mathbf{P_k} + \mathbf{v}\mathbf{v}^T \right| \tag{29}$$

which is equivalent to the following problem by matrix determinant lemma [26]:

$$\mathbf{v}_{k+1} = \underset{\mathbf{v} \in \mathcal{U} - \mathcal{V}_k}{\operatorname{argmax}} \mathbf{v}^T \mathbf{P_k}^{-1} \mathbf{v} \tag{30}$$

Similar to the linear algorithm described in Sect. 3.2, the inverse of $\mathbf{P_k}$ can be updated as following:

$$\mathbf{P}_{\mathbf{k+1}}^{-1} = \mathbf{P_k}^{-1} - \frac{\mathbf{P_k}^{-1} \mathbf{v}_{k+1} \mathbf{v}_{k+1}^T \mathbf{P_k}^{-1}}{1 + \mathbf{v}_{k+1}^T \mathbf{P_k}^{-1} \mathbf{v}_{k+1}} \tag{31}$$

As can be seen, the optimal method of nonlinear NPDOD is essentially the same as that of linear NPDOD. The only difference is that the data points $\mathbf{x}_i (i = 1, \ldots, m)$ are replaced by $\mathbf{u}_i (i = 1, \ldots, m)$.

5 Experiment

In this section, we evaluate the performance of our proposed algorithm for terrain classification. We compare our algorithm with AOD [6] and TED [7] on a toy example in Sect. 5.1. In Sect. 5.2, we describe the data set and feature extraction used in our experiments. The comparative experimental results are present in Sect. 5.3.

5.1 A toy example

A toy example is given in Fig. 1. The data contain two circles with random noise added. There are twenty points on the big circle while ten points on the small circle. We apply AOD, TED, and our proposed NPDOD to select the most informative points on the data set. Here, $\text{SVM}_{\text{active}}$ cannot be applied due to the lack of labeled points. Fig. 1 shows that the points selected by NPDOD can indeed reflect the manifold structure of the data set. Besides, the points selected by our NPDOD algorithm can better represent the data set while both AOD and TED select the points from the big circle. Even though the points selected by AOD or TED are labeled, we are still unable to perform classification since all of labeled points belong to the same class.

5.2 Experimental settings

In this subsection, we describe the experimental settings. The points selected by active learning or random algorithm are used as the training data to train a classifier, and the unselected points are used as the testing data. The classification accuracy of different training data is used to measure the performance of each algorithm.

To demonstrate how our proposed algorithm improves the classification performance, we compare the following algorithms:

- *Random sampling* method, which randomly selects points from the data set as training data.
- *Neighborhood preserving regression* (NPR) which is a semi-supervised learning algorithm
- *Laplacian regularized D-optimal design* (LapRDD) which is an active learning algorithm combining experimental design and graph Laplacian.
- *Neighborhood preserving D-optimal design* (NPDOD) proposed in this paper.

Random sampling and NPDOD are active learning algorithms while NPR is a semi-supervised learning algorithm which uses both labeled and unlabeled points. For NPR, the user is required to label several points which are

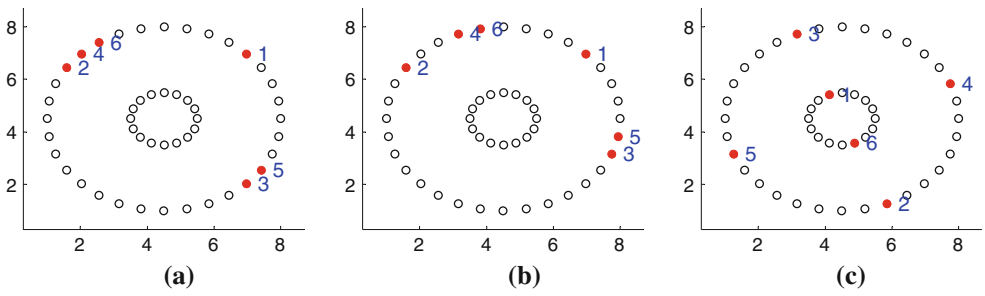


Fig. 1 Data selection by different active learning algorithms. The *numbers* beside the *selected points* denote the orders they were selected. Obviously, the points selected by NPDOD can better represent the original data set. **a** AOD, **b** TED, **c** NPDOD

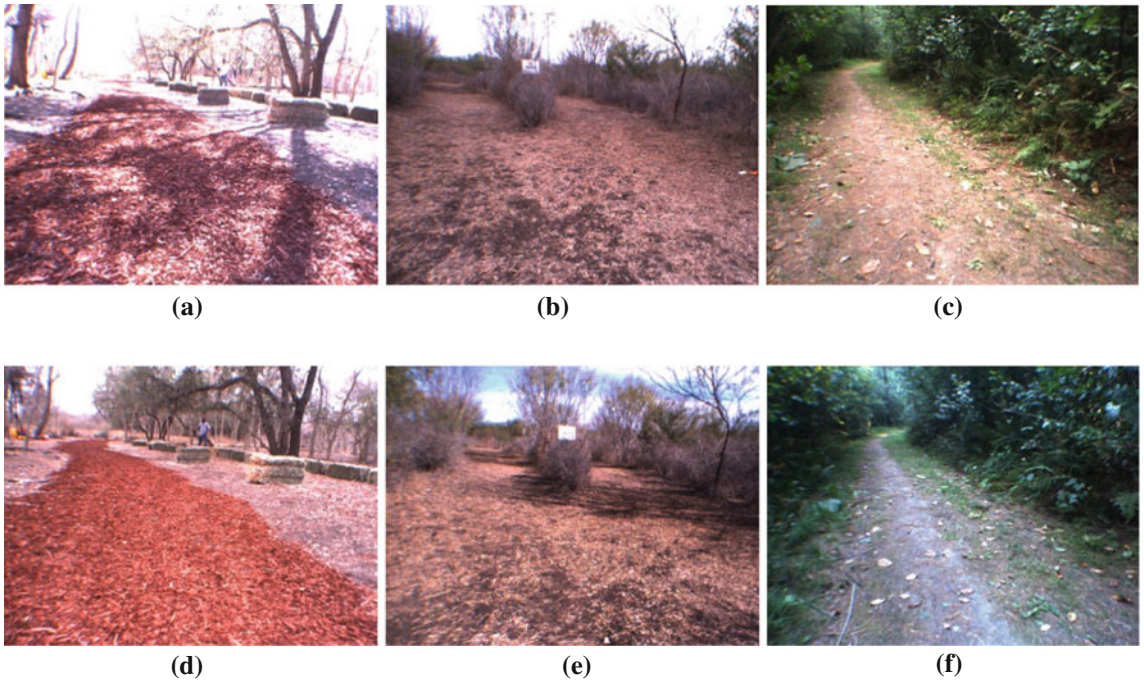


Fig. 2 Representative images from each of the six data sets. **a** Data set 1A (DS1A), **b** Data set 2A (DS2A), **c** Data set 3A (DS3A), **d** Data set 1B (DS1B), **e** Data set 2B (DS2B), **f** Data set 3B (DS3B)

selected randomly, whereas for LapRDD and NPDOD, the points are selected by the algorithm itself. It would be important to note that the SVM classifier is used in random sampling, LapRDD and NPDOD algorithm for classification.

There are three parameters in our algorithm. The number of nearest neighbors (p) is set to be 10 while the parameters λ_1 and λ_2 are empirically set to be 1e−5 and 1e−3, respectively. Linear kernel is used in proposed method and SVM classifier.

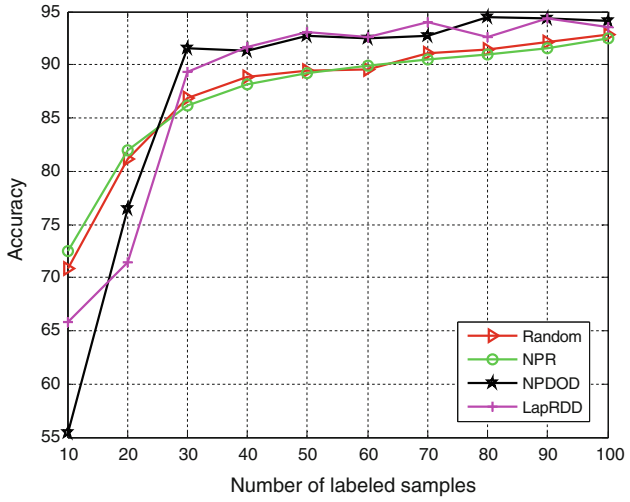
The natural data sets [27, 28] used here are taken from logged field tests conducted by DARPA evaluators. Overall, three scenarios are considered. Each scenario is associated with two distinct image sequences, each representing a different lighting condition. There are thus six data sets total. Representative images are shown in Fig. 2. Each data set consists of a 100-frame, hand-labeled image sequence. Each image was manually labeled, with each pixel being placed into one of three classes: OBSTACLE, GROUNDPLANE, or UNKNOWN. The data sets, hand labelings, and a tool to aid in labeling are all available on the Internet [29].

In experiments, feature extraction method is fixed as color histogram [30]. To create a color histogram, color intensities in each of the three color channels (R, G, and B) in the neighborhood of the reference pixel are binned. The number of bins b (here, fixed at five) and the window dimension $c_w \times c_h$ (fixed at 16 × 16) are parameters of the color histogram feature extraction techniques. Using three color channels and five bins per channel results in a feature image with feature depth d of 15 values (three channels × five bins per channel).

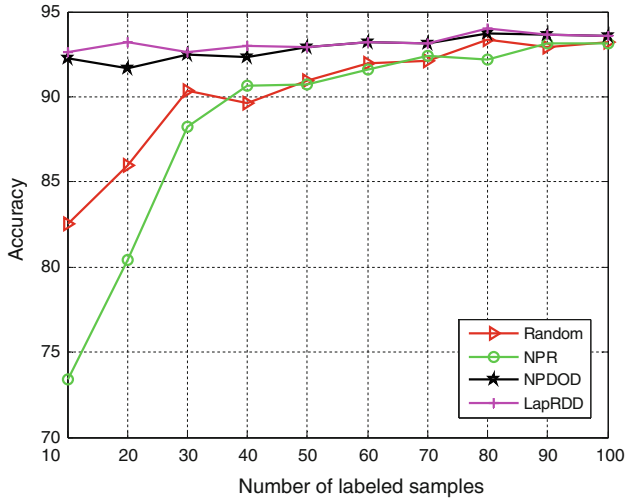
5.3 Experimental results

We use the data sets DS1A, DS2A, and DS3A in this experiment. For each data set, 20 images are extracted. We extract 40 points from each image and the number of OBSTACLE and GROUNDPLANE is the same. The points of UNKNOWN are all ignored. So, there are 800 samples for each data set. We then apply each active learning algorithm to select $k = (10, 20,\ldots,100)$ samples for labeling. The labeled samples are used as training samples while the rest unlabeled samples are used as testing samples.

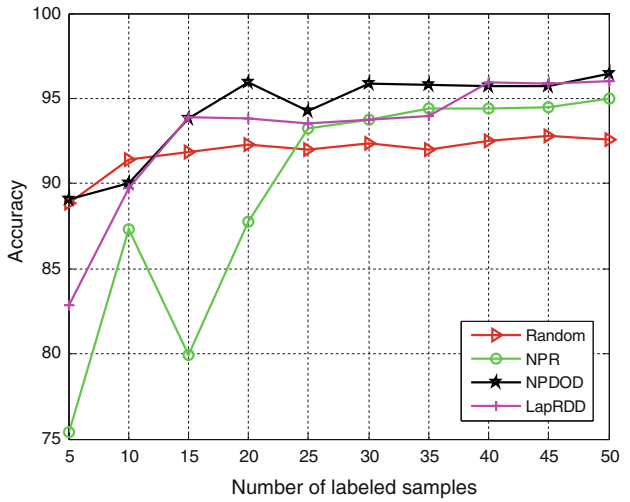
The classification accuracy is shown in Fig. 3. Our NPDOD algorithm outperforms both Random and NPR algorithm in most case in the three datasets. Especially in DS1A and DS2A, the classification accuracy obtained by using 60 samples selected by proposed NPDOD algorithms is comparable to those by using 100 samples selected by Random and NPR algorithms. Compared with LapRDD, NPDOD achieves a better result in DS1A and DS3A. Only in DS2A, LapRDD gets a little higher accuracy than NPDOD. LapRDD is comparable to NPDOD since it is also based on optimal experimental design and combines graph Laplacian regularized regression. However, proposed algorithm NPDOD performs best as a whole.



(a)



(b)



(c)

Fig. 3 Classification results on the data set of DS1A, DS2A, and DS3A. The samples selected by the active learning algorithm are used as training data and the unselected samples are used as testing data. **a** The classification accuracy on DS1A, **b** The classification accuracy on DS2A, **c** The classification accuracy on DS3A

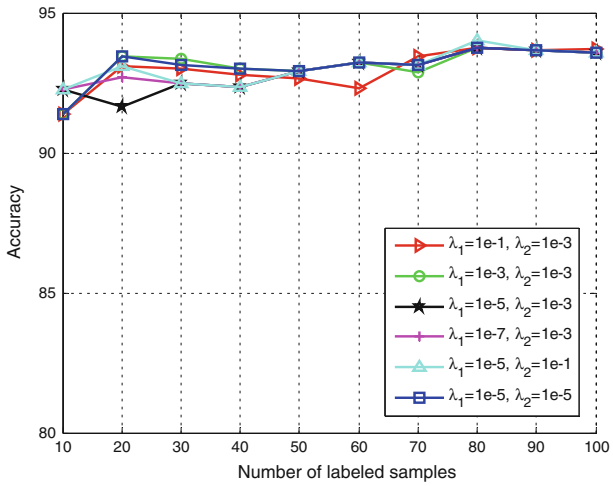


Fig. 4 Classification results on the data set of DS2A with different parameters

In order to prove that the proposed algorithm is robust to the parameters' values, we perform experiments with different values of λ_1 and λ_2 on DS2A .

Figure 4 shows how the performance of NPDOD varies with parameters λ_1 and λ_2. As we can see in Fig. 4, proposed algorithm performance is not sensitive to parameters values. So it is a robust method which can achieve a stable and encouraging result.

6 Conclusion

In this paper, we have introduced a novel active learning algorithm called NPDOD. Our algorithm is motivated from neighborhood preserving regression and OED. For each sample, it may be represented as a linear combination of its p nearest neighbors, a natural assumption is that the label of this sample can also be computed as a linear combination of the labels of its p nearest neighbors. We select the most representative samples such that the global reconstruction error is minimized. Experimental results on terrain classification show the effectiveness of proposed approach.

Central to the proposed algorithm is neighborhood preserving based on the locally linear reconstruction. The reconstruction coefficients are computed by the idea of LLE [24]. But the disadvantage is that the p nearest

neighbor search is computationally expensive. In this work, we adopt D-optimal criterion to minimize the determinant of covariance matrix. However, it remains unclear how other criteria work under this framework. Moreover, the parameter selection is an interesting problem to research which is especially important and difficult for active learning.

Acknowledgments This project is supported by NSFC of China (Grants No. 60632050, No. 60705006, No. 60873151, and No. 60973098).

References

1. Zhu X (2005) Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin Madison
2. Zhou D, Bousquet O, Lal TN et al (2004) Learning with local and global consistency. Adv Neural Inf Process Syst 16:321–328
3. Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 7:2399–2434
4. Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. J Artif Intell Res 4:129–145
5. Settles B (2009) Active learning literature survey, computer sciences. Technical Report 1648, University of Wisconsin Madison
6. Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs with SAS. Oxford University Press, Oxford
7. Yu K, Bi J, Tresp V (2006) Active learning via transductive experimental design. In: Presented at the 23rd international conference of machine learning, Pittsburgh, PA
8. Yu K, Zhu S, Xu W, Gong Y (2008) Non-greedy active learning for text categorization using convex transductive experimental design. InSIGIR'08: proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA, pp 635–642
9. Tong S, Koller D (2002) Support vector machine active learning with applications to text classification. J Mach Learn Res 2:45–66
10. Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: MULTIMEDIA'01: proceedings of the ninth ACM international conference on Multimedia. ACM, New York, NY, USA, pp 107–118
11. Goh K-S, Chang EY, Lai W-C (2004) Multimodal concept-dependent active learning for image retrieval. In: Presented at the ACM conference multimedia, New York
12. Schohn G, Cohn D (2000) Less is more: active learning with support vector machines. In: Presented at the 17th international conference machine learning, Stanford, CA
13. He X (2010) Laplacian regularized D-optimal design for active learning and its application to image retrieval. IEEE Trans Image Process 19(1):254–263
14. Chen C, Chen Z, Bu J, Wang C, Zhang L, Zhang C (2010). G-Optimal design with laplacian regularization. In: Proceedings of the twenty-fourth AAAI conference on artificial intelligence (AAAI-10)
15. Zhang L, Chen C, Chen W, Bu J, Cai D, He X (2009) Convex experimental design using manifold structure for image retrieval. In: Proceedings of the 17th ACM international conference, New York, USA
16. He X, Ji M, Bao H (2009) A unified active and semi-supervised learning framework for image compression. Comput Vis Pattern Recogn
17. Zhang L, Chen C, Bu J, Cai D, He X, Huang TS (2011) Active learning based on locally linear reconstruction. IEEE Trans Pattern Anal Mach Intell 33(10):2026–2038
18. Shen J, Ju B, Jiang T et al (2011) Column subset selection for active learning in image classification. Neurocomputing 74:3785–3792
19. Lu K, Zhao J (2011) Neighborhood preserving regression for image retrieval. Neurocomputing 74:1467–1473
20. Flaherty P, Jordan MI, Arkin AP (2005) Robust design of biological experiments. In: Presented at the advances in neural information processing systems 18, Vancouver, BC, Canada
21. Chung FRK (1997) Spectral graph theory, Regional Conference Series in Mathematics, vol 92
22. Wan M, Lai Z, Jin Z (2011) Feature extraction using two-dimensional local graph embedding based on maximum margin criterion. Appl Math Comput (AMC) 217(23):9659–9668
23. Wan M, Lai Z, Shao J, Jin Z (2009) Two-dimensional local graph embedding discriminant analysis (2DLGEDA) with its application to face and palm biometrics. Neurocomputing (IJON) 73(1–3):197–203
24. Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
25. He X, Cai D, Yan S, Zhang H-J (2005) Neighborhood preserving embedding. In: IEEE international conference on computer vision, Beijing, China, pp 1208–1213
26. Harville DA (1997) Matrix algebra from a statistician's perspective. Springer, New York
27. Procopio MJ, Mulligan J, Grudic G (2009) Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments. J Field Robot 26(2): 145–175
28. Procopio (2007) An experimental analysis of classifier ensembles for learning drifting concepts over time in autonomous outdoor robot navigation. University of Colorado
29. Procopio MJ Hand-labeled DARPA LAGR data sets (2010) Available at http://www.mikeprocopio.com/labeledlagrdata.html
30. Stockman G, Shapiro LG (2001) Computer vision. Prentice Hall, Upper Saddle River