Neurocomputing 329 (2019) 495–505

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Active learning with simultaneous subject and variable selections

Yuan-chin Ivan Chang[a,*], Ray-Bing Chen[b]

[a] Academia Sinica, Taipei, Taiwan
[b] National Cheng Kung University, Tainan, Taiwan

ARTICLE INFO

Article history:
Received 14 January 2018
Revised 8 June 2018
Accepted 7 November 2018
Available online 15 November 2018

Communicated by Dacheng Tao

MSC:
62J05
68T05
68Q32

Keywords:
Active learning
Classification
Effective variable
Stopping rule

ABSTRACT

Training data are essential for learning classification models. Therefore, if only a limited number of labeled subjects are available for use as training samples, whereas a considerable amount of unlabeled data already exists, then it is always desirable enlarging the training set by labeling more subjects in order to ameliorate classification models. When it is costly in time and capital to label unlabeled subjects, it is crucial to know how many labeled subjects are necessary for training a satisfactory classification model. Although, active learning methods can gradually recruit new unlabeled subjects and disclose their label information to enlarge the size of the training set, there is a lack of discussion about the size of training samples in the literature. Hence, when/how to appropriately stop an active learning procedure is studied in this paper. Since the sequential subject recruiting strategy is used in active learning procedures, it is natural to adopt the idea of sequential analysis to dynamically and adaptively determine the training sample size for learning. In this study, we propose a stopping criterion for a linear model-based active learning procedure, such that this learning process will asymptotically achieve its best possible empirical performance, in terms of the area under receiver the operating characteristic curve (ROC), when the procedure is stopped. Other statistical properties of the proposed procedure, including estimation consistency and variable selection, are also studied. The numerical results using both synthesized and a real example are reported.

1. Introduction

The quantity and quality of training data are essential for learning a good classification rule. When the labeled data are limited and a bunch of unlabeled data is available, how to use these unlabeled samples effectively and efficiently becomes a challenging problem. It happens often in modern data analysis scenarios; for example, when we want to utilize an existing data set, collected for other purposes, for a new research goal, some necessary information, such as label information of subjects for constructing a new classification rule, may only exist implicitly in this data set. Thus, if a classification rule based on this data set is of interest and labeling each of them is costly in time and capital, then determining how many labeled samples are required and which samples should be labeled first for constructing a satisfactory classifier is a crucial problem. Active learning methods are popularly discussed this situation. In general, when an active learning method is used, the unlabeled data are sequentially annotated and recruited into a training set based on the current model information, such that we can learn the target model more economically [1–3].

Active learning methods have been used in different applications including automatic cell segmentation [4], multimedia annotation and retrieval [5], synthetic aperture radar image classification [6]; meta learning [7], meta-cognitive machine learning [8], computer aided medical diagnosis [9] and so on. Practitioners usually use different subject selection strategies according to their application needs. In [10], authors conducted some empirical comparisons of different learning strategies in different scenarios. Some popularly discussed strategies in the literature include statistical experimental design methods [11], transductive experimental design methods [12,13], graph structure-based methods [13,14] and expert opinion-based methods [6,15]. Recently, the prediction uncertainty [16], the total budget cost [17] and modern penalized regression methods [18] are also considered in some active learning algorithms.

This adaptive subject selection feature makes active learning procedures naturally sequential procedures with adaptive samples. Thus, how and when to stop an active learning procedure is of interest. In [19], authors applied an entropy-based stopping

* Corresponding author.
E-mail addresses: ycchang@sinica.edu.tw, yc.ivan.chang@me.com (Y.-c. Ivan Chang), rbchen@mail.ncku.edu.tw (R.-B. Chen).

https://doi.org/10.1016/j.neucom.2018.11.036
0925-2312/

criterion to an active learning procedure, which requires considering the prediction probabilities of all unlabeled subject and is computationally intensive. In [20], authors discussed some stopping criterions and all of them require some complicated evaluation of the unlabeled subjects and, therefore, are usually time-consuming. As far as we know, the truncated-type, pre-fixed training sizes are usually used in practice. However, this truncated-type sample sizes has little information about the model performance and cannot ensure the quality of the final model. Therefore, how to have a stopping criterion, which only uses the information of the learning samples and can also ensure us a satisfactory final performance, is an important problem in active learning.

Due to the subject recruiting steps in active learning methods, new samples are adaptively selected into the training set. Thus, the samples used for training in this kind of procedures are no longer statistically independent. Therefore, we adopt the idea of the stopping time in sequential analysis to dynamically determine the training sample size of active learning procedures. In particular, we will treat it as a sequential estimation problem of a stochastic regression model with adaptive explanatory variables. The subject selection process may depend on the training target of a classifier [2,21,22]. The area under an ROC curve (AUC) is a common performance target in classification. In [23], authors also studied active learning procedures that maximize AUC. Here, we use a stochastic regression model as a base model and show that the proposed stopping criterion based on this model can achieve the maximum possible area under ROC curve, asymptotically. For more information about stochastic regression, please refer to [24]. Moreover, the effective variables of the classification model can also be identified sequentially through the learning procedure. The key features of the our algorithm include:

- **Classification model** Using a stochastic linear regression model for binary classification problems;
- **Subject selection** Applying both the *D*-optimality experimental design and the uncertainty criteria in the subject selection stage;
- **Variable selection** Simultaneously determine the effective variables for the final model with an adaptive shrinkage estimating method;
- **Stopping criterion** Adopting a stopping criterion such that our algorithm will asymptotically achieve the maximum area under ROC curve, when the learning procedure is stopped.

The rest of this paper is organized as follows. We will first review the use of linear regression models in binary classification scenarios and compare the classification performance of a linear regression model to with the Fisher linear discriminant analysis (LDA). We then state the propose algorithm and its statistical properties. Both synthesized and real data sets are used to illustrate our method.

2. Method

Despite the fitting errors of applying a linear model to a binary response data set, it has been shown that, for binary classification problems, the regression coefficient estimates of a linear model is related to the Fisher LDA direction for a given training data set. The proof of it can be readily found in many machine learning textbooks such as, for example the work of [25]. For convenience, we will give a summary of the result in this section following a brief description of Fisher's linear discrimination analysis (LDA). In addition, because the proposed stopping criterion mainly depends on the classification model and can be used with different subjects selection strategy, we will first discuss the properties of the proposed stopping criterion and then describe the subject selection method used in this paper after that.

2.1. Binary classification with a linear model

Suppose that there are two classes, say C_1 and C_2. Let $P_1 = Pr(C_1)$ and $P_2 = Pr(C_2)$, $P_1 + P_2 = 1$, be the proportions of C_1 and C_2 in the population, respectively. Let $\mathbf{x} \in R^p$ be the random vector recording the measures of a subject. Assume that $\mathbf{x}$ in Class C_i, $i = 1, 2$, follows a normal distribution with mean μ_i with a covariance matrix Σ. Define the vector $\boldsymbol{\ell} \in R^p$ as the projection direction of the Fisher's LDA, and let $\ell_0 \in R$ signify a cutting point. Then the Fisher's LDA will assign $\mathbf{x}$ to Class 1 if $\boldsymbol{\ell}^T\mathbf{x} + \ell_0 > 0$; otherwise assign $\mathbf{x}$ to C_2, where

$$\boldsymbol{\ell} = \Sigma^{-1}(\mu_2 - \mu_1) \tag{1}$$

$$\ell_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2) - \log\left(\frac{P_2}{P_1}\right). \tag{2}$$

It is known that the Fisher's linear discriminant analysis has the maximum AUC under normality assumption. In addition, since AUC is scale-invariant and independent of cutting point, it implies that the vector $L = c \cdot \boldsymbol{\ell}$, for any constant $c > 0$, will have the same AUC as that of the Fisher's LDA. Hence, if the estimate of a coefficient vector of a linear model is parallel to this Fisher's LDA direction, then it will also achieve the same AUC.

Let $\{t_1, t_2\} \in R$ be the labels of two classes, that is,

$$t_i = \begin{cases} t_1 \text{ for all } \mathbf{x}_i \in C_1; \\ t_2 \text{ for all } \mathbf{x}_i \in C_2, \end{cases} \tag{3}$$

and let $n = n_1 + n_2$ be the size of the training set, where n_1 and n_2 are the sample sizes of C_1 and C_2, respectively. Define $\mathbf{z}_j^T = (1, \mathbf{x}_j)$, for $j = 1, \ldots, n$, be the augmented measurement vectors for subject j. For $i = 1, 2$, let $\mathbf{u}_i$ be a vector of n_i 1s, and let the matrix $\mathbf{X}_i$ be a matrix with n_i rows, containing the training set patterns in C_i, and p columns. Define a matrix

$$\mathbf{Z_n} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{X}_1 \\ \mathbf{u}_2 & \mathbf{X}_2 \end{bmatrix}. \tag{4}$$

Thus, the least squares solution a set of samples of size n, $\tilde{\boldsymbol{\beta}}_n$, that minimizes $\|\mathbf{Z}_n\boldsymbol{\beta} - \mathbf{Y}_n\|^2$ satisfies the equation

$$\mathbf{Z}_n^T\mathbf{Z}_n\boldsymbol{\beta}_n = \mathbf{Z}_n\mathbf{Y}_n,$$

where $\mathbf{Y}_n = (y_1, \ldots, y_n)^T$ with $y_j = t_1$, for $j = 1, \ldots, n_1$ and $y_j = t_2$, for $j = n_1 + 1, \ldots, n_1 + n_2$. That is, $\mathbf{Y}_n$ is a vector with first n_1 elements equal to t_1 and the rest n_2 elements equal to t_2, and the linear model defined above is based on the augmented measurement vectors $\mathbf{z}_j \in R^{p+1}$.

Assume that $t_1 \neq t_2$, and let $\tilde{\boldsymbol{\beta}}_n^T = (\tilde{\beta}_{0,n}, \tilde{\boldsymbol{\beta}}_{1,n}^T)$, we have that

$$\tilde{\boldsymbol{\beta}}_{1,n} = \frac{\alpha}{n}S_n^{-1}(\mathbf{m}_1 - \mathbf{m}_2), \tag{5}$$

$$\tilde{\beta}_{0,n} = -\frac{1}{n}(n_1\mathbf{m}_1 + n_2\mathbf{m}_2)^T\tilde{\boldsymbol{\beta}}_{1,n} + \frac{n_1}{n}t_1 + \frac{n_2}{n}t_2, \tag{6}$$

where α is a constant, $\mathbf{m}_i$ is the sample mean of the rows of $\mathbf{X}_i$ and

$$S_n \equiv n^{-1}(\mathbf{X_1}^T\mathbf{X_1} + \mathbf{X_2}^T\mathbf{X_2} - n_1\mathbf{m}_1\mathbf{m}_1^T - n_2\mathbf{m}_2\mathbf{m}_2^T)$$

is a sample covariance matrix. The $\tilde{\beta}_{0,n} \in R$ in (6) is the empirical cutting point for the given data. If we ignore the scale parameter α in (5), then the vector $\tilde{\boldsymbol{\beta}}_{1,n}$ has a similar form to that in (1) by replacing Σ as S_n. Hence, $\tilde{\boldsymbol{\beta}}_{1,n}$ is parallel to the "Fisher's LDA direction" $\boldsymbol{\ell}$ as in (1). (Please see [25, Chapter 5.2.4, page 231 for

further details], and also note that for the convenience of the following discussion, Eq. (4) here is different from theirs, because the different notations are used.)

We use linear models for classification purposes, thus only the classification/prediction performance is of interest. Since it is usually not advisable to fit binary response data with linear models, linear models act as "fake" models here from the model fitting perspectives. Hence, from a classification viewpoint, we want the coefficient estimates obtained from this fake linear model to be as close to the direction of Fisher's LDA as possible. Thus, the conventional sequential procedure for the constructing a fixed size confidence ellipsoid for the regression parameters $\boldsymbol{\beta}$, such as that suggested by [24], can be used. Below, we prove that if the proposed stopping criterion is used, then when the active learning procedure is stopped, the angle between the estimates of the regression coefficient vector, $\tilde{\boldsymbol{\beta}}_{1,n}$, and the projection vector of the Fisher linear discrimination analysis, $\boldsymbol{\ell}$, will be close to zero. Therefore, the AUC obtained based on the coefficient-estimate-based stopping criterion will also converge to the best AUC that a linear model can achieve with the given training data set.

2.2. Stochastic linear regression and active learning

Suppose that $\{\mathbf{x}_j \in R^p,\ j = 1, \ldots, n\}$ is a set of samples of size n, and let t_i, $i = 1, 2$, be the class labels as defined before. Define σ-field $\mathcal{F}_n = \sigma\{(y_j, \mathbf{z}_j), j = 1, 2, \ldots, n\}$ for $n \geq 1$ and $\mathcal{F}_0 = \sigma\{\emptyset, \Omega\}$. Assume that random observations $\{(y_j, \mathbf{z}_j), i = 1, 2, \ldots\}$ satisfy that $E[y_j|\mathcal{F}_{j-1}] = \mathbf{z}_j^T\boldsymbol{\beta}_0$ for all $j \geq 1$. Define $e_j = y_j - E[y_j|\mathcal{F}_{j-1}]$ for each $i \geq 1$. It follows that $\{e_n\colon\ n \geq 1\}$ is a sequence of martingale differences with respect to σ-fields $\mathcal{F}_n$. Let $\tilde{\boldsymbol{\beta}}_n = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{z}_i^T\boldsymbol{\beta})^2$ be the ordinary least squares estimate (LSE) of a random sample of size n. Assume that

(A1) $\lim_{n\to\infty} \sum_{j=1}^n \mathbf{z}_j\mathbf{z}_j^T/n$ exists and is a positive definite.

(A2) $\sup E(|e_n|^\alpha|\mathcal{F}_n) < \infty$ almost surely for some $\alpha > 2$,

(A3) The maximum and minimum eigenvalues of matrix $\sum_{j=1}^n \mathbf{z}_j\mathbf{z}_j^T$, $\lambda_{\max}(n)$ and $\lambda_{\min}(n)$, satisfy $\lambda_{\min}(n) \to \infty$ and $\log(\lambda_{\max}(n)) = o(\lambda_{\min}(n))$ with probability one.

Then following [24], it is shown that $\tilde{\boldsymbol{\beta}}_n$ converges to the true parameters, $\boldsymbol{\beta}_0$, with probability one.

Let $\lambda \equiv \lambda(n)$ be a non-random function of n such that for some $0 < \delta < 1/2$ and $\gamma > 0$,

$$n^{1/2}\lambda \to 0 \text{ and } n^{1/2+\gamma\delta}\lambda \longrightarrow \infty, \text{ as } n \to \infty. \tag{7}$$

Define $I_0 = diag(I(\beta_{01} \neq 0), \ldots, I(\beta_{0p} \neq 0))$ as a $p \times p$ diagonal matrix; that is, I_0 is the true (unknown) indicator matrix of the nonzero "effective" variables. Let $\lambda_j = \lambda|\tilde{\beta}_{nj}|^{-\gamma}$, the empirical variable indicator $I_{nj}(\epsilon) = I(\sqrt{n}\lambda_j < \epsilon)$, where $j = 1, \ldots, p$, and

$$I_n(\epsilon) = \text{diag}\{I_{n1}(\epsilon), \ldots, I_{np}(\epsilon)\}$$

be a $p \times p$ diagonal matrix. By the law of the iterated logarithm, we have that $n^{1/2-\eta}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = o(1)$ almost surely as $n \to \infty$ for some $\eta > 0$. Thus, define $\hat{\boldsymbol{\beta}}_n = I_n(\epsilon)\tilde{\boldsymbol{\beta}}_n$ as an asymptotic shrinkage estimate (ASE) for $\boldsymbol{\beta_0}$ [see also 26]. Let p_0 denote the number of effective variables and let $\hat{p}_0 \equiv \hat{p}_0(n) \equiv \sum_{i=1}^p I_{nj}(\epsilon)$ as an estimator of p_0. The notation $AUC_{\mathbf{v}}$ denotes the AUC with the projection vector $\mathbf{v}$. Then we have the following theorem.

Theorem 2.1. *Assume that (A1) to (A3) are satisfied. Then $\hat{p}_0 \to p_0$ with probability one, and $\lim_{n\to\infty} E[\hat{p}_0] = p_0$. Suppose that $N(t)$ is a positive integer-valued random variable for which $N(t)/t$ converges to 1 in probability as $t \to \infty$. Then with probability one, as t goes to infinity, $\hat{\boldsymbol{\beta}}_{N(t)} \equiv (\hat{\beta}_{0,N(t)}, \hat{\boldsymbol{\beta}}_{1,N(t)}^T)^T \to \boldsymbol{\beta}_0 \equiv (\beta_0, \boldsymbol{\beta}_1^T)^T$ and*

$$AUC_{\hat{\boldsymbol{\beta}}_{1,N(t)}} \to AUC_{\boldsymbol{\ell}}.$$

Moreover,

$$\sqrt{N(t)}(\hat{\boldsymbol{\beta}}_{N(t)} - \boldsymbol{\beta}_0) \longrightarrow N(0, \sigma^2 l_0 \Sigma^{-1} I_0) \quad \text{in distribution as } t \to \infty. \tag{8}$$

From previous discussion, we know that $\boldsymbol{\beta}_1$ is parallel to $\boldsymbol{\ell}$. It implies that $AUC_{\boldsymbol{\beta}_1}$ is equal to AUC_{ℓ}, because AUC is scale-invariant. Thus, if $\hat{\boldsymbol{\beta}}_{N(t)} \to \boldsymbol{\beta}_0$ almost surely as $t \to \infty$, then the AUC of the classifier with a direction of $\hat{\boldsymbol{\beta}}_{1,N(t)}$ will converge to the AUC of the classifier with the vector $\boldsymbol{\beta}_1$ with probability one (please refer to [26] for the detailed proof about the consistency of $\hat{\boldsymbol{\beta}}_{N(t)}$).

2.2.1. Stopping criterion and training sample size

Ideally, we desire a pre-determined sample size that can ensure that the classifier constructed via an active procedure within this size of training set can perform as good as the classifier trained with the whole data set with all label information. However, in practice, it is difficult to determine how well a classifier can perform since there is a lack of label information for all subjects, and there is no known relation between training size and performance. Moreover, due to the adaptive subject selection nature of active learning procedures, it is hard to have a pre-specified sample size for a given prediction goal. The usual cross-validation methods are not useful in this situation [19]. Truncating a learning procedure at a pre-determined sample size is a commonly used method; however, such a truncated sample size cannot guarantee the best usage of the available data. Some stopping criterions, based on entropy and the idea of uncertainty, were suggested in the literature; these methods usually require the application of an extra evaluation scheme to the remain unlabeled subjects in the data set at each learning stage and whether the learning process is stopped will depend on this evaluation. Hence, these type of methods are computationally intensive, and impractical when the size of data set is large [19,20].

Here we follow the sequential nature of active learning procedures and apply a stopping criterion based on the ideas in sequential analysis, which depends only on the subjects in the current training set. The stopping criterion that is proposed below is targeted at AUC. When the sampling is stopped, the performance of the proposed classification rule will approach to the optimal AUC of the model with the given data set with a prespecified probability. The stopping criterion for the linear model-based classifier is stated below and its properties are summarized as Theorem 2.2 below.

Let $\hat{\sigma}_n^2 = (n-p)^{-1}\mathbf{Y}_n^T(I_n - \mathbf{Z}_n^T(\mathbf{Z}_n\mathbf{Z}_n^T)^{-1}\mathbf{Z}_n)\mathbf{Y}_n$. Let a_k^2 be a $1-\alpha$ quantile of the chi-square distribution with $\hat{p}_0(n)$ degrees of freedom, where $\hat{p}_0(n)$ is the estimate of the number of effective variables as before. Assume further that

(A4) there exists a non-random positive definite symmetric matrix B_n and a continuously increasing function $\rho(\cdot)$ such that

$$B_n^{-1}(\sum_{j=1}^n \mathbf{z}_j\mathbf{z}_j^T)^{1/2} \to I_{p+1} \text{ and } \max_{1\leq j\leq n} ||B_n^{-1}\mathbf{z}_j|| \to 0 \text{ in probability}, \tag{9}$$

$$\sum_{j=1}^n \mathbf{z}_j\mathbf{z}_j^T/\rho(n) \to \Sigma \text{ almost surely,}$$

where Σ is a positive definite matrix. (10)

(A5) $\lim_{n\to\infty} \log(v_{max}(n))/n = 0$ almost surely.

Eq. (10) states a condition for $\rho(n)$ as an increasing function of sample size for general stochastic linear regression cases, which will depend on the sequence of the random design $\mathbf{z}_i$'s. In active

learning procedures, we usually have a specified subject selection rule in advance. For example, if the *D*-optimality method is used, then $\rho(n) = O(n)$ is sufficient.

Let

$$R_n = \left\{ \mathbf{z} \in R^p : \frac{S_n}{\rho(n)} \le \frac{d^2}{\nu_n} \text{ and } z_j = 0 \text{ for } I_{nj(\epsilon)} = 0,\ 1 \le j \le p \right\}, \tag{11}$$

where ν_n is the maximum eigenvalue of $\rho(n) I_n(\epsilon)(\mathbf{Z}_n \mathbf{Z}_n^T)^{-1} I_n(\epsilon)$ and $\rho(n)$ is a continuously increasing function satisfies Assumption (A4). See also [24] for further discussion about sequential confidence ellipsoids. Let d be the half width of the maximum axis of the confidence ellipsoid R_n and define a stopping rule

$$T = \inf\left\{ n \ge n_0 : (\hat{\sigma}_n^2 + n^{-1}) \le d^2 n / (a_n^2 \nu_n) \right\}. \tag{12}$$

Let $\theta(u, v)$ denote the angle between vectors u and v. Then we have following results.

Theorem 2.2. *Assume (A1) to (A5) are satisfied. Let the stopping time T be defined as in (12), and $\hat{\theta}_T = \theta(\hat{\boldsymbol{\beta}}_{1,T}, \boldsymbol{\ell})$ be the angle between $\hat{\boldsymbol{\beta}}_{1,T}$ and the LDA direction $\boldsymbol{\ell}$, then*

(i) $\lim_{d\to 0} \theta(\hat{\boldsymbol{\beta}}_{1,T}, \boldsymbol{\ell}) = 0$ *with probability one, and*

$$\lim_{d\to 0} P\left[0 \le \hat{\theta}_T \le \cos^{-1}\left(\frac{\|\boldsymbol{\ell}\| - d}{\|\boldsymbol{\ell}\| + d} \right) \right] \ge 1 - \alpha,$$

(ii) $\mathrm{AUC}_{\hat{\boldsymbol{\beta}}_{1,T}} \to \mathrm{AUC}_{\boldsymbol{\ell}}$ *almost surely as* $d \to 0$.

Theorem 2.2 (i) means that the angle between the estimated projection direction and the Fisher's LDA direction is eventually less than $\cos^{-1}((\|\boldsymbol{\ell}\| - d)/(\|\boldsymbol{\ell}\| + d))$ with probability $1 - \alpha$. Theorem 2.2 (ii) says that the empirical AUC of the proposed active learning procedure will converge to its theoretical optimal AUC, asymptotically.

Proof of Theorem 2.2:

Once a subject is selected and a label is assigned, then the conventional regression theorems can be applied to this training set. It can be shown that if the *D*-optimal selection scheme is used as a subject selection criterion, the (A2), (A3) and (A5) are satisfied. Assumption (A4) is a stability assumption on the design matrix, which is required for applying the martingale central limit theorem. (For further discussion regarding the existence of a nonrandom matrix B_n, please refer to [24] and the references therein.)

From the definition of stopping time, we know that T goes to infinity as d goes to 0 with probability one. Hence, following similar arguments in [26], it can be shown that $\hat{\boldsymbol{\beta}}_{1,T} \to \boldsymbol{\beta}_1$ almost surely. This implies that $\mathrm{AUC}_{\hat{\boldsymbol{\beta}}_{1,T}}$ eventually converges to $\mathrm{AUC}_{\boldsymbol{\ell}}$ as $d \to 0$. Similarly, it is known that $P(\boldsymbol{\beta}_0 \in R_T) \to 1 - \alpha$ as $d \to 0$. Moreover, by definition, the maximum axis of R_n is no greater than $2d$. Let $\hat{\boldsymbol{\ell}}$ denote the estimated direction $\hat{\boldsymbol{\beta}}_{1,T}$. Then, with simple vector algebra, it is shown that as $d \to 0$, with probability no less than $1 - \alpha$, that

$$\cos(\hat{\theta}) = \frac{< \hat{\boldsymbol{\ell}}, \boldsymbol{\ell} >}{\|\hat{\boldsymbol{\ell}}\| \|\boldsymbol{\ell}\|} \ge \left(\frac{\|\boldsymbol{\ell}\| - d}{\|\boldsymbol{\ell}\| + d} \right)^2.$$

It is clear that if d goes to 0, then $\cos(\hat{\theta})$ goes to 1, which implies that $\hat{\theta}$ converges to 0. It implies that with probability no less than $1 - \alpha$, if d is small enough, then

$$0 \le \hat{\theta} \le \cos^{-1}\left(\left(\frac{\|\boldsymbol{\ell}\| - d}{\|\boldsymbol{\ell}\| + d} \right)^2 \right),$$

□

To apply the results above, certain regularity conditions are required on the newly selected subjects. No specific selection scheme is specified, and other design schemes may also be used here. The stopping time mainly depends on the classification model and here a stochastic linear model is used. In the following subsection, we will discuss how the *D*-optimal design can be used here for demonstration purposes.

2.2.2. Subject selection

When constructing a classification rule, the availability of label information is the major difference between active learning procedures and conventional classification training procedures. Once the all subjects are selected and labeled, the computing algorithm of these two procedures is the same. That is, how subjects are selected without label information is the key issue.

Following the definition of $\mathbf{Z}_n$ in (4), we have that

$$\mathbf{Z}_n^T \mathbf{Z}_n = \begin{bmatrix} \mathbf{u}_1^T \mathbf{u}_1 + \mathbf{u}_2^T \mathbf{u}_2 & \mathbf{u}_1^T \mathbf{X}_1 + \mathbf{u}_2^T \mathbf{X}_2 \\ (\mathbf{u}_1^T \mathbf{X}_1 + \mathbf{u}_2^T \mathbf{X}_2)^T & \mathbf{X}_1^T \mathbf{X}_1 + \mathbf{X}_2^T \mathbf{X}_2 \end{bmatrix}. \tag{13}$$

(Note that for simplicity, the subscript n of $\mathbf{Z}_n$ will be omitted when there is no ambiguity.) It is known that $n = \mathbf{u}_1^T \mathbf{u}_1 + \mathbf{u}_2^T \mathbf{u}_2$, and $\mathbf{X}_1^T \mathbf{X}_1 + \mathbf{X}_2^T \mathbf{X}_2 = \sum_{i=j}^n \mathbf{x}_j \mathbf{x}_j^T$. Thus, if matrix $\sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T$ is positive definite, then it implies that $\mathbf{Z}^T \mathbf{Z}$ is also positive definite. Hence, if the new subjects are selected according *D*-optimal criterion to maximize the determinant of $\sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T$, then Assumptions (A1) and (A3) are satisfied. Moreover, because t_1, t_2 are bounded, (A2) is satisfied. In this case, the selection procedure will not depend on the label information of $\mathbf{x}$. Hence, by Theorem 2.1, it is shown that $AUC_{\hat{\boldsymbol{\beta}}_{1,n}} \to AUC_{\boldsymbol{\ell}}$ almost surely as $\min(n_1, n_2) \to \infty$.

Because $\hat{\boldsymbol{\beta}}_{1,n}$ is an estimate of the LDA direction based on the current observed samples of size n, we define $r(x) = \hat{\boldsymbol{\beta}}_{1,n}^T \mathbf{x}$ as a score function of sample with measurement vector $\mathbf{x}$. For a given cutting point $\hat{\beta}_0$, the distance between the risk score of a subject $\mathbf{x}$ is equal to $r(\mathbf{x}) - \hat{\beta}_0$. The unlabeled subjects are then listed based on $|r(\mathbf{x}) - \hat{\beta}_0|$ in increasing order. For given constant $r \in (0, 1)$, based on the principle of uncertainty, we will search the subject within the top $r \times 100\%$ based on the *D*-optimal criterion. That is, within this searching range defined by r, the subject with measurement vector $\mathbf{x}_{n+1}$ that maximizes the determinant of matrix $\sum_{j=1}^{n+1} \mathbf{x}_j \mathbf{x}_j^T = \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T$ will be recruited into the training set after being labeled. Note that during this selection procedure, the observations up to n-th stage, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, will not be altered. This feature is useful when the size of unlabeled data is very large, since we can limit the search range. The performance of classifiers may be diminished if we only search for a very small range of subjects. Hence, to achieve the same performance the process with a smaller search range may take longer learning time, which can be seen from the numerical results later.

Remark. Note that although the stopping criterion proposed here relies mainly on the classification model, the design part still has some effect on the efficiency of the learning procedures. That is, the sample size used may vary as the design methods and subject selection schemes. Moreover, the computational cost of the statistical experimental design also depends on the model. When a nonlinear model, such as a logistic model, is used, the optimal design may depend on the unknown regression parameters. Hence, some iterative algorithms and the idea of local optimality will be applied. [27] used Bayesian design as an alternative. However, it may be computational stable in the early learning stage when the size of labeled subjects is small. It is hard to apply to data sets with lengthy explanatory variables. Here we take the advantages of the connection between a linear model and Fisher's LAD, and the ease of experimental design under linear models as well, such that the computational cost is diminished and the final classification performance is under control.

Table 1
Performance of Active Learning procedures: with and without a variable selection feature.

	Testing data		Training sample size	
Variable selection	Accuracy	AUC	Positive	Negative
with Var. Sel.	0.8210 (0.0120)	0.9038 (0.0093)	492.3 (33.51)	491.7 (33.91)
without Var. Sel.	0.8196 (0.0109)	0.8196 (0.0109)	609.7 (23.36)	609.4 (22.78)

The means of the accuracy and AUC for testing data sets with and without the variable selection features are presented. In addition, the positive and negative subjects used in the training stage are also reported. The numbers within the parentheses are their corresponding standard deviations.

3. Numerical results

We apply the proposed active learning algorithm to some synthesized data and two real data sets. For the synthesized data, the theoretical value of AUC is known and will be used as the baseline for comparison. We also report the results of active learning algorithms without variable selection feature for comparison purposes. For the real data sets used here, their true labels are actually available, and these label information of each subject will be revealed only when it is recruited into the training set. For illustration purposes, we will pretend that only a limited number of them are labeled in the beginning as the initial training set, and the most of them are unlabeled. That is, the label information in the original data will be used as domain experts during the learning process in our numerical studies. Once a subject is selected into the training set, then we know its label.

3.1. Synthesized data

Assume that the data of Group 0 (the negative group) are generated from a normal distribution with a mean vector $\mu_0 = (1, 0.5, -0.4, -0.9, 0, \ldots, 0)$ in R^{20}, and a covariance matrix Σ below:

$$\Sigma = \begin{pmatrix} A & O \\ O & B \end{pmatrix},$$

where

$$A = \begin{pmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0.2 & 0 \\ 0 & 0.2 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{pmatrix},$$

O is a matrix of zeros and B is a 16×16 identity matrix. That is, those dummy variables are uncorrelated to the effective variables, and have a mean vector equal to 0 and a unit variance-covariance matrix. The data of Group 1 (the positive group) is generated with a mean vector $\mu_1 = (1, -0.5, 0.4, 0.9, 0, \ldots, 0)$ in R^{20} and the same variance-covariance matrix. It follows that the optimal direction based on Fisher's LDA is $\ell = (0, -0.7511, 0.6335, 1.2, 0, \ldots, 0)$ in R^{20}, with the corresponding theoretical $AUC = 0.9044$ under this simulation setup.

The numerical results reported here are based on 1000 replications. In each simulation run, 1500 training data from each group are generated, and 500 data are generated from each group as the testing data set. The initial sample size used is equal to 40. Table 1 summarizes the means of the accuracy and AUC for the testing data set, and the sample sizes selected from both positive and negative groups in its training stage. Note that although there are a total of 3000 samples that are generated as a training set in each run, only a small part of them is selected during the active learning procedures.

As defined before, the constant d denotes the half width of the maximum axis of R_n, and $r \in (0, 1)$ denotes the search range. When $r = 0.1$, we will search the top 10% subjects around the cutting point, and $r = 1$ means that the algorithm will search all of the remaining unlabeled subjects in each stage. The proposed sequential procedure, with $d = 0.1$ and $r = 0.2$, selects 492 samples, on average, from each class for training in these simulation studies. From Table 1, we can see that the proposed linear model-based active learning procedure with the proposed stopping criterion has an average AUC equal to 0.9038 (standard deviation equal to 0.0093), which is very close to its theoretical optimal value of 0.9044. Fig. 1 is a box-plot of the sample sizes used for two groups, and it shows that around 1000 labeled subjects are selected, on average, for training. Fig. 2 shows that the estimated coefficient vector and the LDA direction are very close; the most of the angles between the estimated vector and the theoretical LDA direction are less than 6 degrees. Fig. 3 reveals the similar information. It is clearly shown in Fig. 3 that only the nonzero components in the LDA vector ℓ are selected, and the other variables are only selected randomly with very low frequencies. In other words, the selection procedure can select the correct variables in this simulation study. Here, we choose $t_1 = 0$ and $t_2 = 1$ in this numerical study. Note that the simulation studies with different values of ts are also conducted and their results are similar.

For comparison purposes, we conduct the conventional active learning algorithm using a linear model, under the same parameters ($r = 0.2$ and $d = 0.1$) and stopping criterion, without the shrinkage estimating procedure for selecting variables, and also summarize the results in Table 1. Because the stopping criterion depends on the coefficient estimates of the regression model, the procedure without variable selection will have a lengthy coefficient vector to be estimated and therefore will require a larger training sample size. Because the accuracy depends on the cutting point selection for the classification scores, the averages testing accuracy of the procedure without variable selection can still reach a similar accuracy to that of the proposed algorithm. However, without a variable selection step, this model will falsely include some unwanted variables, which diminish its AUCs. Table 1 clearly shows that the average testing AUC without a variable selection feature is significantly smaller than that of the algorithm with a variable selection feature.

3.2. Real data I: polycystic ovary syndrome and androgen excess data

This dataset was originally collected at Taipei Medical University Wan Fang Hospital for an infertility study. There are 27 measures recorded for each subject including age, menarche, systolic, diastolic, BMI, total testosterone, insulin, mFG score (FGS), and so on. From the literature, the infertility rate is highly associated with polycystic ovaries. It is also known that the polycystic ovary syndrome (PCOS) is the most common androgen-excess (AE) disorder, and affects between 5% and 10% of all women. (The details about PCOSAE definition and diagnosis criteria can be found in [28].) Thus, medical doctors would like to know whether we can have a diagnostic/classification rule using this existing data set and to find out which variables among these 27 measures are also highly associated with PCOSAE. To this end, some medical experts re-examined all cases and then labeled all the subjects as PCOSAE and non-PCOSAE. There are 462 subjects and, among them, 206 subjects are diagnosed as PCOSAE patients by medical experts. In each

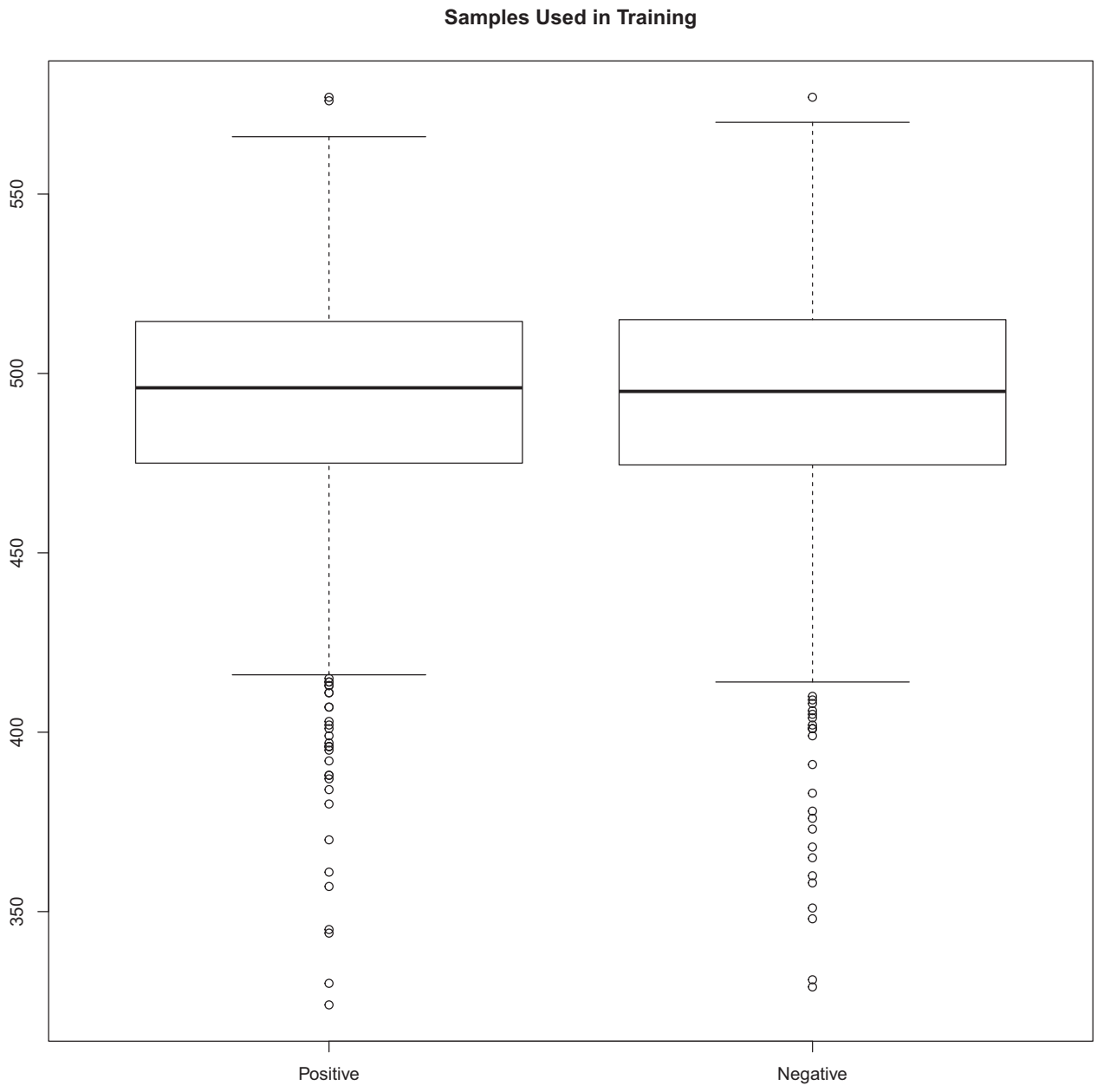

Fig. 1. Sample size boxplot.

run, we randomly select 50 subjects from both the PCOSAE group and non-PCOSAE groups as an independent testing set, which will not be considered as training samples during the learning procedure. We then apply a logistic regression model to all 362 training subjects in each run using their true labels and all 27 measures together. The average prediction AUC of logistic models obtained based on 1000 replications is equal to 0.862, with a standard deviation equal to 0.025. Below, we use the same dataset to demonstrate how the proposed active learning can be applied to this situation without re-examining the record of each subject in order to label each of them for the new problem of interest; that is, to find out the possible diagnostic rule for detecting the status of the PCOSAE using these 27 measures of the previous infertility study.

For comparison purposes, we also used the randomly select 50 subjects from both the PCOSAE group and non-PCOSAE group as an independent testing set in each run. We then start our active learning procedure with only 60 labeled subjects – 30 subjects in each group, and select one subject at a time from the rest of the subjects in the data set until the proposed stopping criterion is fulfilled as discussed before. Note that the label information about the selected subjects is not revealed until they are included in training. That is, no label information is used during the subject selection stage. This procedure will be repeated 1000 times as in the simulation study, and the results reported here are based on 1000 replications. We expect that with the proposed active learning procedure, we can train a classification rule that can have a satisfactory classification performance in terms of AUC with only a smaller size of the labeled subjects. Because there is no the distributional information known for these variables, the corresponding theoretical/optimal AUCs cannot be available. We will use the AUCs of logistic models in the analysis described before as a baseline for comparison purposes. The sample sizes difference between active learning procedures and the corresponding total size is the size of samples, which remain “unlabeled” (or unrevealed)

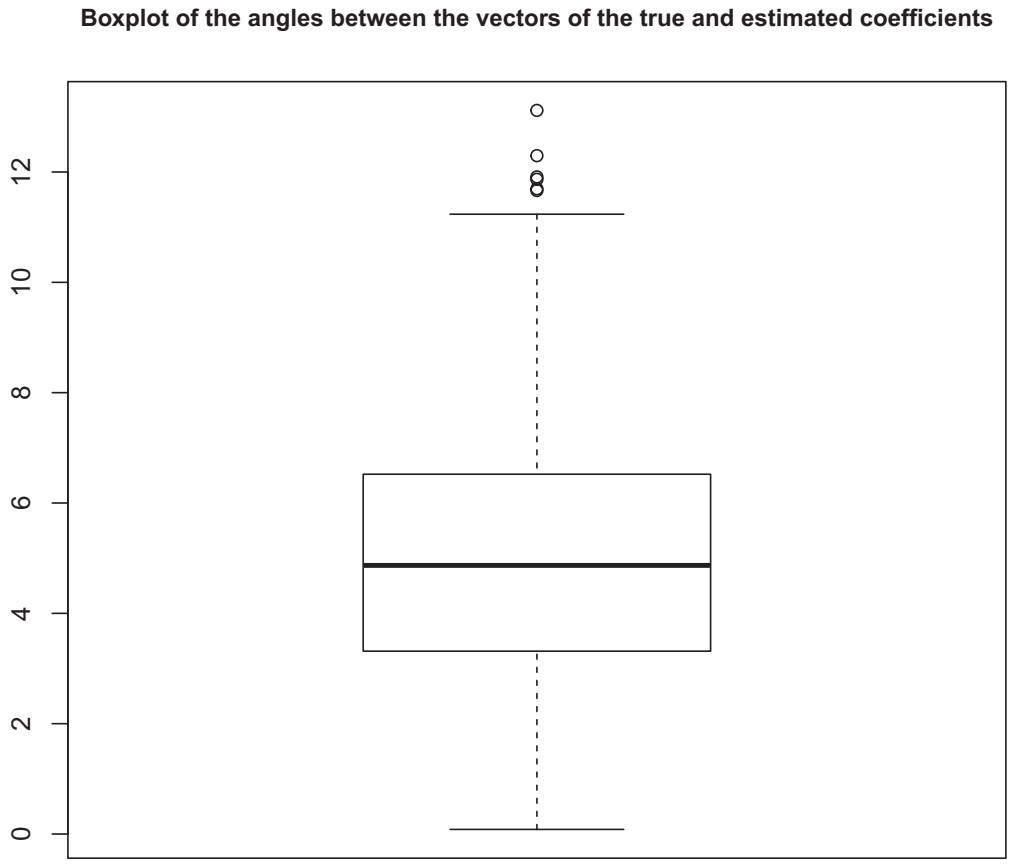

Fig. 2. Box-plot of absolute values of angles between the estimated direction and the LDA direction.

Table 2
PCOS: accuracy and area under curve.

Range/d	Accuracy	Area under curve
r = 1.0/d = 0.15	0.742 (0.061)	0.819 (0.063)
r = 1.0/d = 0.20	0.732 (0.062)	0.813 (0.064)
r = 1.0/d = 0.30	0.726 (0.065)	0.805 (0.069)
r = 0.5/d = 0.15	0.746 (0.059)	0.825 (0.058)
r = 0.5/d = 0.20	0.730 (0.063)	0.811 (0.062)
r = 0.5/d = 0.30	0.723 (0.063)	0.803 (0.067)
r = 0.1/d = 0.15	0.734 (0.059)	0.815 (0.058)
r = 0.1/d = 0.20	0.725 (0.059)	0.807 (0.062)
r = 0.1/d = 0.30	0.718 (0.061)	0.799 (0.067)

Table 3
PCOS: sample sizes used.

Range/d	Positive group	Negative group	Total used
r = 1.0/d = 0.15	139.10 (24.24)	177.72 (34.81)	316.82 (58.40)
r = 1.0/d = 0.20	114.08 (35.30)	140.99 (49.27)	255.07 (84.03)
r = 1.0/d = 0.30	91.32 (42.44)	109.94 (59.06)	201.26 (101.13)
r = 0.5/d = 0.15	139.15 (25.46)	172.20 (36.41)	311.35 (61.23)
r = 0.5/d = 0.20	112.11 (35.51)	131.08 (50.24)	243.19 (85.10)
r = 0.5/d = 0.30	85.85 (43.16)	98.94 (58.71)	184.79 (101.46)
r = 0.1/d = 0.15	130.62 (31.21)	152.59 (43.35)	283.21 (73.91)
r = 0.1/d = 0.20	104.09 (42.31)	115.43 (59.41)	219.53 (101.19)
r = 0.1/d = 0.30	80.75 (46.66)	90.68 (60.59)	171.43 (106.91)

during the active learning procedures. These differences represent the efficiency of using the proposed stopping criterions in these active learning procedures.

Table 2 summarizes the AUCs and accuracies of different confidence ellipsoid criteria d with different searching range r. Note that those numbers inside the parentheses are their corresponding standard deviations. When $d = 0.15$ without any confinement on the searching range, (i.e. $r = 1$) the AUC of the proposed method achieves 0.819 with standard deviation equal to 0.063, which is close to the AUC of the logistic procedure mentioned above.

Table 3 shows that the sample sizes used in two groups for different length values d, and different searching ranges r. For example, if $d = 0.2$ is used and there is no limitation of the searching range, then the proposed method has an AUC equal to 0.813 with a standard deviation equal to 0.064, while the required total sample size is 255 on average, which is only slightly more than half of the sample size used in the previous logistic model. That is, the proposed method achieves 94.3% (i.e. 0.813/0.862) AUC performance with 61% (i.e. 255/412) samples of used in the logistic model.

The coefficient estimates under different simulation setups can be found in Appendix B. Since there is no known true model for this PCOS reported in the literature, we only summarize the coefficient estimates and their corresponding standard deviations.

3.3. Real data II: MAGIC gamma telescope data set

The MAGIC gamma telescope data set was generated by a Monte Carlo program, Corsika, described in [29]. This data set is available at the UCI Machine Learning Repository web site [30]. It includes 12332 gamma (signal) and 6688 hadron (background) samples, for a total of 19020 samples, with 10 variables and a label (signal or background) variable for each observation. This data set is often treated as a binary classification problem, as described on the web site. Because incorrectly classifying a background event as signal is worse than classifying a signal event as background,

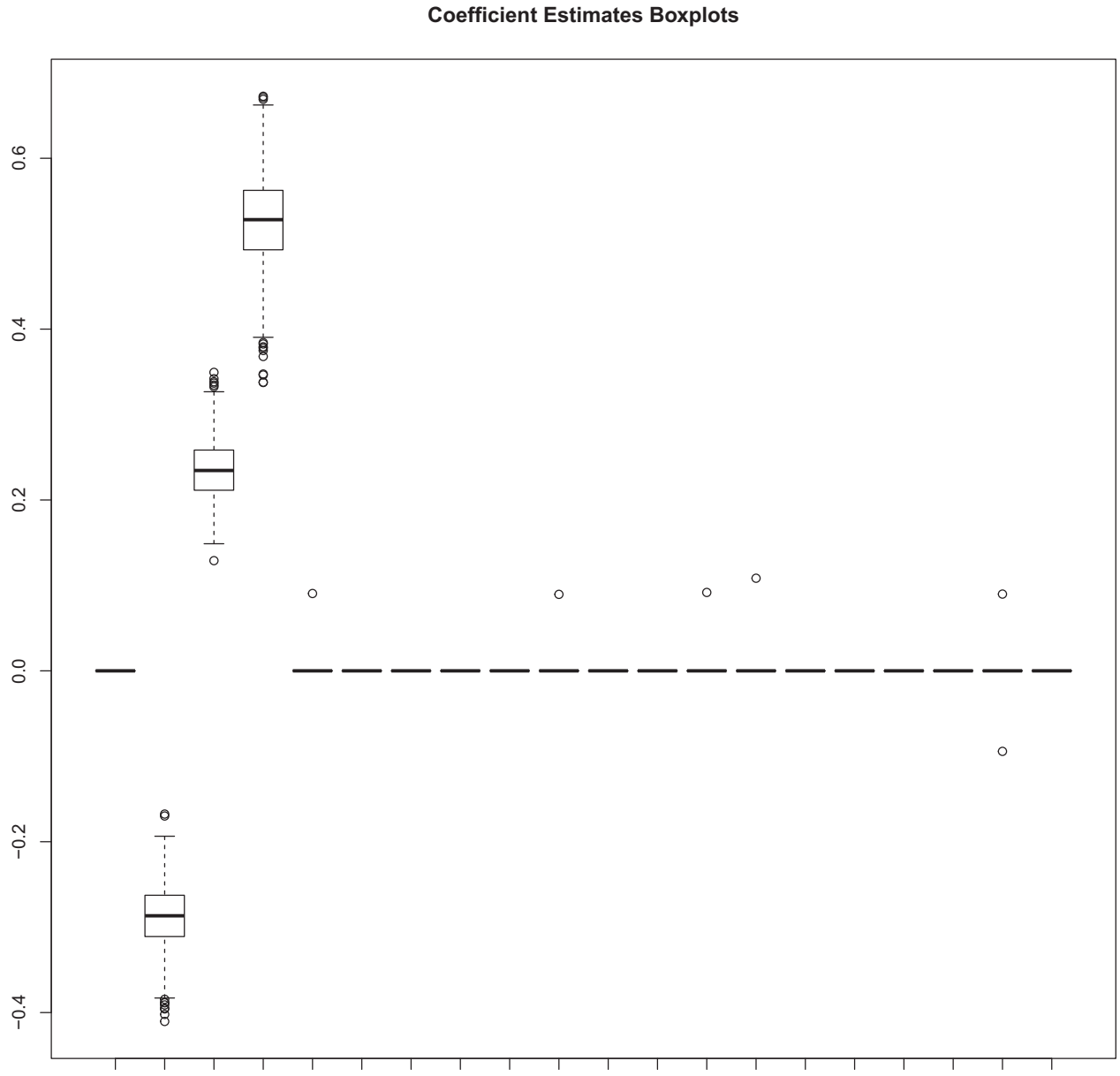

Fig. 3. The boxplot of the coefficient estimates of variables.

the simple measure of classification accuracy is not meaningful for these data; instead, ROC-curve-related measures of classification performance are recommended. For each run in our study, we randomly selected 600 signal observations and 400 background observations to form the testing set, and the remaining observations were used as the training set. Then, 40 initial training samples were randomly selected from the training set in each run. (We will refer to this data set as MAGIC gamma in this paper.)

It is known that when the confidence ellipsoid criteria d is smaller, then we will require larger training sample size to achieve the required precision. We use slightly larger d than that in the PCOS case to save the computational time. Table 4 summarizes the AUCs and accuracies of different confidence ellipsoid criteria d with different searching range r. Table 5 shows that the sample sizes used in two groups for different length values d, and different searching ranges r. We also summarize the coefficient estimates and variable selection results under different setup in Appendix B. Because d is related to the precision of estimates, the procedure with smaller d will require a larger training data size. We can see from Table 4 that for the same d, the AUCs of the cases with larger r are better than the cases with smaller r. When r is large, an active learning procedure can effectively select useful informative samples from a larger range to improve its performance at a cost

Table 4
Accuracy and area under curve for magic gamma data set.

Range/d	Accuracy		Area under curve	
r = 1.0/d = 0.3	0.7107	(0.0549)	0.7567	(0.0708)
r = 1.0/d = 0.5	0.6994	(0.0615)	0.7445	(0.0787)
r = 1.0/d = 1.0	0.6925	(0.0624)	0.7350	(0.0820)
r = 0.5/d = 0.3	0.6758	(0.0783)	0.7167	(0.0944)
r = 0.5/d = 0.5	0.6810	(0.0736)	0.7213	(0.0901)
r = 0.5/d = 1.0	0.6860	(0.0649)	0.7259	(0.0836)
r = 0.1/d = 0.3	0.6569	(0.0865)	0.6946	(0.1028)
r = 0.1/d = 0.5	0.6686	(0.0770)	0.7078	(0.0935)
r = 0.1/d = 1.0	0.6864	(0.0656)	0.7280	(0.0828)

Table 5
Sample sizes used for magic gamma data set.

Range/d	Positive group	Negative group	Total used
r = 1.0/d = 0.3	94.92 (65.53)	73.80 (46.32)	168.72 (110.89)
r = 1.0/d = 0.5	89.22 (51.30)	68.24 (30.34)	157.47 (81.13)
r = 1.0/d = 1.0	76.61 (39.03)	60.42 (21.97)	137.03 (60.54)
r = 0.5/d = 0.3	117.79 (87.50)	98.69 (74.07)	216.49 (159.48)
r = 0.5/d = 0.5	93.08 (50.87)	77.46 (39.90)	170.54 (88.52)
r = 0.5/d = 1.0	72.39 (36.92)	59.27 (21.59)	131.65 (57.70)
r = 0.1/d = 0.3	129.72 (106.03)	123.42 (114.70)	253.15 (213.37)
r = 0.1/d = 0.5	93.47 (57.36)	82.03 (49.45)	175.49 (103.07)
r = 0.1/d = 1.0	73.75 (37.01)	60.26 (21.53)	134.01 (57.73)

of searching time from a larger range. That is also the reason why for a fixed d, the total sample size used of the case with smaller r is greater than that of the case with larger r, since there are usually less candidate samples to be selected and then a larger sample size is required to make the classification model stable. The choice of parameters r and d may depend on the size of data set, the variable length, available computation power and applications, and therefore the choices of them will reply on the practitioners and can be viewed as turning parameters. Although there is no gold standard for these two parameters, to start with larger d and smaller r may help us to efficiently understand the data in hands and find a suitable pair of parameters.

4. Conclusion

From both capital and time-efficiency considerations, it is common that people want to dig out some useful information from an existing data set. For example, it is often that there may be a lack of the essential label information required for constructing a classification rule for the new problem since the available dataset might be collected for different purposes. In this case, it is highly desirable to have a method/procedure that requires only to label a part of the selected subjects from this existing dataset, instead of each of them, and active learning method is a kind of machine learning methods or sequential statistical procedures that can be used in this scenario. We propose a stopping criterion for a linear model-based active learning procedure for binary classification problems in this paper. Using the advantages of stochastic regression and the asymptotic shrinkage estimate method, we are able to adaptively select subjects and to simultaneously determine the effective variables. The proposed stopping criterion based on the estimate of the confidence ellipsoid for the regression parameter which guarantees that the final risk score projection direction converges to the Fisher's LDA direction such that the AUC of the proposed method also approaches the AUC of the Fisher's LDA when the learning (sampling) process is stopped. In this paper, the D-optimal criterion in the experimental design is used as a searching criterion and we also borrow the idea of the uncertainty sampling to confine the search range for new training subjects, which can reduce the searching time when the size of the candidate subjects is large. Here, we show that the AUC of the Fisher's LDA can be achieved by using a linear model with a D-optimal criterion. However, the sample size used in this case cannot be as small as that in the usual sequential estimation, when the labels of subjects are known in advance, which is reasonable due to the lack of label information for recruiting new subjects. The proposed stopping criterion is based on the idea of confidence ellipsoid estimation. This kind of stopping criterions can be applied to active learning procedures that use different design methods as their searching criterions, as long as the selected subjects satisfy the assumptions of our theorems here. However, the properties of classifier, such as efficiency, may vary depending on the selection criterions. In general, the computational cost for selecting new subjects of the linear model-based methods, as the proposed method, is lower than that of the nonlinear model-based methods, because the experimental design to select the next explored point is usually computationally intensive in nonlinear model cases [see 27,31]. This is another advantage of using linear models here.

Searching the best subjects in active learning procedures is always the most time-consuming part, and this situation becomes worse when the size of the unlabeled subjects is huge as in the "big data analysis" scenarios. This difficulty suggests that applying a clustering algorithm to the unlabeled data before conducting an active learning procedure may reduce the searching time and can be a promising approach. Since many clustering algorithms are reported in the literature, the way in which to match a specific active learning algorithm to a particular clustering method is an important problem that will be studied and reported elsewhere.

Appendix A. Least squares solution and Fisher's LDA direction

Please note that the arguments here are similar to that in [25], with only slight modification in order to simplify the notations used in stochastic regression.

Following the notations defined before, it implies that the least squares solution $\hat{\boldsymbol{\beta}}$ must satisfy $\mathbf{Z}_n^T\mathbf{Z}_n\hat{\boldsymbol{\beta}} = \mathbf{Z}_n^T\mathbf{Y}$. It implies that

$$\begin{bmatrix} \mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2 & \mathbf{u}_1^T\mathbf{X}_1 + \mathbf{u}_2^T\mathbf{X}_2 \\ (\mathbf{u}_1^T\mathbf{X}_1 + \mathbf{u}_2^T\mathbf{X}_2)^T & \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_2^T\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_{0,n} \\ \hat{\boldsymbol{\beta}}_{1,n} \end{bmatrix} = \begin{bmatrix} n_1 t_1 + n_2 t_2 \\ n_1 t_1 \mathbf{m}_1 + n_2 t_2 \mathbf{m}_2 \end{bmatrix}, \tag{A.1}$$

where $\mathbf{m}_i$ is the sample mean of the rows of $\mathbf{X}_i$, for $i = 1, 2$. Hence,

$$\hat{\beta}_{0,n} = -\frac{1}{n}(n_1\mathbf{m}_1 + n_2\mathbf{m}_2)^T\hat{\boldsymbol{\beta}}_{1,n} + \frac{n_1}{n}t_1 + \frac{n_2}{n}t_2 \tag{A.2}$$

$$(n_1\mathbf{m}_1 + n_2\mathbf{m}_2)\hat{\beta}_{0,n} + (\mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_2^T\mathbf{X}_2)\hat{\boldsymbol{\beta}}_{1,n} = n_1 t_1\mathbf{m}_1 + n_2 t_2\mathbf{m}_2 \tag{A.3}$$

Assume that $t_1 - t_2 \neq 0$, then it implies that

$$\left[nS_n + \frac{n_1 n_2}{n}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1^T - \mathbf{m}_2^T)\right]\hat{\boldsymbol{\beta}}_{1,n} = \frac{n_1 n_2}{n}(t_1 - t_2)(\mathbf{m}_1 - \mathbf{m}_2) \tag{A.4}$$

where

$$S_n \equiv n^{-1}(\mathbf{X_1}^T\mathbf{X_1} + \mathbf{X_2}^T\mathbf{X_2} - n_1\mathbf{m}_1\mathbf{m}_1^T - n_2\mathbf{m}_2\mathbf{m}_2^T)$$

is the sample covariance matrix as defined before. Since $(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1^T - \mathbf{m}_2^T)$ is in the direction of $\mathbf{m}_1 - \mathbf{m}_2$, it implies that

$$\hat{\boldsymbol{\beta}}_{1,n} = \frac{\alpha}{n}S_n^{-1}(\mathbf{m}_1 - \mathbf{m}_2),$$

where α is a non-zero constant. That is, the solution above is an empirical estimate of the Fisher's LDA direction.

Appendix B. Coefficient estimates

The tables in this section summarize the coefficient estimates and their standard deviations under different ds and rs. Since the ASE method is used for the parameter estimate, we can see that many of them are shrunk to zero in these tables. It is clear that for binary response data, the model fitting property of a linear model will not be as good as, for example, a logistic model. In fact, we should not apply a linear model to a binary response data set under conventional model-fitting consideration. Please note that here we want to take the computational advantage and the ease of D-optimality in a linear model, and our goal is to use it for classification purposes. Hence, these coefficient estimates should not be interpreted as that in conventional model fitting scenarios. The values of the coefficient estimates of variables for the linear model here just represent the relative impact of each variable to the final classification model.

Table B.1 summarizes the coefficient estimates and their standard deviations of the numerical studies using the PCOS data. For the numerical studies using the Magic gamma data set under different setups, in addition to the coefficient estimates and their corresponding standard deviations when using MAGIC gamma data (Tables B.4–B.6), we also summarize the results of the variable selection frequencies (Table B.7).

Table B.1
PCOS: variable coefficient estimates: range= 1.0.

	d = 0.15	d = 0.2	d = 0.3
Age	−0.0002 (0.0061)	0.0000 (0.0000)	−0.0006 (0.0128)
Menarche	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
Systolic	0.0006 (0.0097)	0.0001 (0.0040)	0.0016 (0.0195)
Diastolic	−0.0003 (0.0058)	−0.0003 (0.0060)	−0.0005 (0.0156)
Height	0.0050 (0.0510)	0.0008 (0.0452)	0.0029 (0.0397)
Weight	0.0275 (0.2611)	0.0384 (0.2297)	0.0193 (0.2064)
BMI	−0.0491 (0.2513)	−0.0491 (0.2221)	−0.0299 (0.1981)
Hip	0.0042 (0.0386)	−0.0008 (0.0296)	0.0012 (0.0305)
WHR	0.0003 (0.0071)	0.0000 (0.0000)	0.0008 (0.0185)
IntY	−0.0222 (0.0555)	−0.0211 (0.0631)	−0.0265 (0.0733)
HOMAC	0.0000 (0.0000)	0.0000 (0.0000)	0.0002 (0.0062)
GOT	−0.0074 (0.0403)	−0.0072 (0.0430)	−0.0039 (0.0334)
GPT	0.0063 (0.0329)	0.0053 (0.0327)	0.0023 (0.0368)
CRP	−0.0006 (0.0105)	−0.0019 (0.0218)	−0.0034 (0.0370)
TSH	0.0000 (0.0000)	0.0000 (0.0000)	−0.0002 (0.0048)
LH	−0.0007 (0.0127)	−0.0001 (0.0112)	0.0001 (0.0220)
FSH	0.0000 (0.0000)	0.0000 (0.0081)	−0.0009 (0.0173)
E2	−0.0108 (0.0506)	−0.0166 (0.0749)	−0.0198 (0.0843)
PRL	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
ToTestC	0.2041 (0.0887)	0.2185 (0.1061)	0.2352 (0.1345)
ADDC	0.0000 (0.0000)	0.0006 (0.0106)	0.0020 (0.0233)
SHBG	−0.0005 (0.0134)	−0.0029 (0.0293)	−0.0066 (0.0515)
FAIC	−0.0009 (0.0123)	−0.0002 (0.0097)	−0.0002 (0.0228)
DHEAS_C	0.0009 (0.0138)	0.0035 (0.0278)	0.0062 (0.0369)
AMHC	0.0000 (0.0000)	0.0003 (0.0097)	0.0000 (0.0000)
X17_OHPC	0.0010 (0.0119)	0.0015 (0.0164)	0.0034 (0.0333)
FGS	0.1148 (0.0917)	0.1050 (0.1047)	0.0931 (0.1120)

Table B.2
PCOS: variable coefficient estimates: range = 0.5.

	d = 0.15	d = 0.2	d = 0.3
Age	−0.0003 (0.0082)	−0.0007 (0.0138)	−0.0012 (0.0163)
Menarche	0.0000 (0.0000)	0.0000 (0.0000)	0.0002 (0.0055)
Systolic	0.0002 (0.0052)	0.0007 (0.0133)	0.0013 (0.0168)
Diastolic	−0.0001 (0.0044)	−0.0004 (0.0071)	−0.0006 (0.0106)
Height	0.0015 (0.0453)	−0.0011 (0.0410)	−0.0014 (0.0371)
Weight	0.0570 (0.2446)	0.0628 (0.2053)	0.0497 (0.1898)
BMI	−0.0673 (0.2361)	−0.0666 (0.2036)	−0.0517 (0.1848)
Hip	−0.0009 (0.0315)	−0.0012 (0.0259)	−0.0005 (0.0271)
WHR	0.0000 (0.0000)	0.0000 (0.0000)	0.0006 (0.0111)
IntY	−0.0214 (0.0559)	−0.0174 (0.0563)	−0.0197 (0.0657)
HOMAC	0.0000 (0.0000)	0.0000 (0.0000)	−0.0002 (0.0073)
GOT	−0.0013 (0.0152)	−0.0025 (0.0237)	−0.0012 (0.0163)
GPT	0.0017 (0.0167)	0.0029 (0.0230)	0.0015 (0.0189)
CRP	−0.0002 (0.0072)	0.0000 (0.0000)	−0.0005 (0.0110)
TSH	0.0000 (0.0000)	0.0000 (0.0000)	−0.0002 (0.0066)
LH	0.0000 (0.0000)	−0.0002 (0.0075)	0.0006 (0.0168)
FSH	0.0000 (0.0000)	0.0000 (0.0000)	0.0006 (0.0175)
E2	−0.0036 (0.0253)	−0.0083 (0.0513)	−0.0123 (0.0688)
PRL	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
ToTestC	0.2132 (0.0865)	0.2286 (0.1084)	0.2493 (0.1435)
ADDC	0.0000 (0.0000)	0.0020 (0.0212)	0.0052 (0.0399)
SHBG	0.0000 (0.0000)	0.0000 (0.0000)	−0.0020 (0.0272)
FAIC	−0.0012 (0.0136)	0.0000 (0.0086)	0.0011 (0.0208)
DHEAS_C	0.0026 (0.0221)	0.0044 (0.0320)	0.0062 (0.0382)
AMHC	0.0000 (0.0000)	0.0002 (0.0076)	0.0004 (0.0080)
X17_OHPC	0.0005 (0.0095)	0.0021 (0.0210)	0.0026 (0.0280)
FGS	0.1078 (0.0905)	0.0818 (0.0979)	0.0757 (0.1054)

Table B.3
PCOS: variable coefficient estimates: range= 0.1.

	d = 0.15	d = 0.2	d = 0.3
Age	0.0000 (0.0000)	−0.0002 (0.0050)	−0.0015 (0.0183)
Menarche	0.0000 (0.0000)	0.0000 (0.0098)	−0.0003 (0.0095)
Systolic	0.0005 (0.0095)	0.0022 (0.0225)	0.0029 (0.0242)
Diastolic	−0.0003 (0.0057)	−0.0004 (0.0070)	−0.0009 (0.0136)
Height	−0.0037 (0.0396)	−0.0047 (0.0342)	−0.0015 (0.0309)
Weight	0.0912 (0.2101)	0.0803 (0.1829)	0.0502 (0.1669)
BMI	−0.0920 (0.2064)	−0.0816 (0.1794)	−0.0497 (0.1612)
Hip	−0.0021 (0.0318)	−0.0006 (0.0245)	−0.0010 (0.0237)
WHR	0.0000 (0.0000)	0.0006 (0.0092)	0.0001 (0.0041)
IntY	−0.0112 (0.0394)	−0.0114 (0.0438)	−0.0164 (0.0599)
HOMAC	0.0000 (0.0000)	0.0000 (0.0000)	−0.0002 (0.0066)
GOT	−0.0009 (0.0118)	−0.0016 (0.0178)	−0.0003 (0.0062)
GPT	0.0011 (0.0146)	0.0026 (0.0205)	0.0013 (0.0160)
CRP	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
TSH	0.0000 (0.0000)	0.0000 (0.0000)	−0.0004 (0.0090)
LH	0.0000 (0.0000)	−0.0001 (0.0041)	0.0003 (0.0132)
FSH	0.0000 (0.0000)	0.0000 (0.0000)	0.0006 (0.0106)
E2	−0.0006 (0.0091)	−0.0012 (0.0132)	−0.0016 (0.0202)
PRL	0.0000 (0.0000)	−0.0004 (0.0087)	−0.0004 (0.0082)
ToTestC	0.2288 (0.0920)	0.2537 (0.1257)	0.2401 (0.1426)
ADDC	0.0010 (0.0156)	0.0012 (0.0173)	0.0014 (0.0175)
SHBG	0.0000 (0.0000)	0.0000 (0.0000)	−0.0005 (0.0098)
FAIC	−0.0006 (0.0091)	−0.0004 (0.0209)	0.0013 (0.0168)
DHEAS_C	0.0016 (0.0175)	0.0035 (0.0287)	0.0058 (0.0367)
AMHC	−0.0001 (0.0043)	0.0000 (0.0000)	0.0014 (0.0176)
X17_OHPC	0.0005 (0.0085)	0.0010 (0.0137)	0.0005 (0.0101)
FGS	0.0737 (0.0844)	0.0587 (0.0825)	0.0609 (0.0942)

Table B.4
Variable coefficient estimates for magic gamma data set: range= 1.

	d = 0.3	d = 0.5	d = 1.0
X28.7967	−0.0280 (0.0694)	−0.0516 (0.1029)	−0.0780 (0.1283)
X16.0021	−0.0018 (0.0181)	−0.0055 (0.0520)	−0.0140 (0.0842)
X2.6449	−0.0028 (0.0473)	−0.0137 (0.0914)	−0.0305 (0.1143)
X0.3918	0.0025 (0.0203)	−0.0013 (0.0466)	−0.0040 (0.1552)
X0.1982	0.0021 (0.0186)	0.0024 (0.0456)	−0.0446 (0.1439)
X27.7004	0.0020 (0.0199)	0.0004 (0.0237)	0.0005 (0.0442)
X22.011	0.0013 (0.0175)	0.0018 (0.0281)	0.0035 (0.0331)
X.8.2027	−0.0003 (0.0126)	−0.0006 (0.0208)	0.0021 (0.0357)
X40.092	−0.1779 (0.0817)	−0.1672 (0.0900)	−0.1571 (0.1004)
X81.8828	−0.0059 (0.0332)	−0.0048 (0.0405)	−0.0014 (0.0318)

Table B.5
Variable coefficient estimates for magic gamma data set: range= 0.5.

	d = 0.3	d = 0.5	d = 1.0
X28.7967	−0.0755 (0.1197)	−0.0850 (0.1394)	−0.0993 (0.1449)
X16.0021	0.0002 (0.0673)	−0.0061 (0.0845)	−0.0146 (0.0975)
X2.6449	−0.0023 (0.1133)	0.0027 (0.1178)	−0.0258 (0.1465)
X0.3918	0.0083 (0.0513)	0.0094 (0.0987)	−0.0198 (0.2706)
X0.1982	0.0115 (0.0470)	0.0145 (0.0939)	−0.0314 (0.2353)
X27.7004	−0.0011 (0.0273)	−0.0017 (0.0356)	−0.0002 (0.0406)
X22.011	0.0055 (0.0400)	0.0066 (0.0446)	0.0074 (0.0473)
X.8.2027	−0.0001 (0.0282)	0.0009 (0.0266)	0.0008 (0.0308)
X40.092	−0.1393 (0.1001)	−0.1520 (0.1011)	−0.1508 (0.1059)
X81.8828	−0.0018 (0.0391)	−0.0004 (0.0291)	−0.0012 (0.0360)

Table B.6
Variable coefficient estimates for magic gamma data set: range = 0.1.

	d = 0.3	d = 0.5	d = 1.0
X28.7967	−0.0974 (0.1504)	−0.0757 (0.1291)	−0.0960 (0.1478)
X16.0021	−0.0206 (0.0996)	−0.0056 (0.0841)	−0.0074 (0.0917)
X2.6449	−0.0356 (0.1422)	−0.0019 (0.1323)	−0.0007 (0.1354)
X0.3918	−0.0185 (0.2753)	0.0158 (0.0779)	0.0145 (0.1151)
X0.1982	−0.0392 (0.2433)	0.0136 (0.0598)	0.0142 (0.1026)
X27.7004	−0.0012 (0.0515)	0.0016 (0.0386)	−0.0012 (0.0451)
X22.011	0.0046 (0.0462)	0.0054 (0.0439)	0.0067 (0.0405)
X.8.2027	−0.0042 (0.0466)	0.0025 (0.0315)	0.0013 (0.0352)
X40.092	−0.1508 (0.1065)	−0.1227 (0.1061)	−0.1404 (0.1041)
X81.8828	−0.0015 (0.0321)	−0.0024 (0.0390)	−0.0007 (0.0326)

Table B.7
Variable Selection Frequencies for Magic Gamma Data Set.

	Range = 1.0			Range = 0.5			Range = 0.1		
	d = 0.3	d = 0.5	d = 1.0	d = 0.3	d = 0.5	d = 1.0	d = 0.3	d = 0.5	d = 1.0
X28.7967	0.147	0.218	0.314	0.317	0.331	0.365	0.311	0.348	0.346
X16.0021	0.010	0.042	0.094	0.076	0.094	0.124	0.094	0.117	0.132
X2.6449	0.058	0.121	0.171	0.175	0.201	0.195	0.226	0.245	0.205
X0.3918	0.016	0.040	0.208	0.044	0.091	0.324	0.096	0.155	0.324
X0.1982	0.013	0.039	0.269	0.066	0.115	0.299	0.082	0.131	0.353
X27.7004	0.010	0.011	0.04	0.021	0.028	0.031	0.030	0.030	0.038
X22.011	0.009	0.014	0.021	0.036	0.044	0.041	0.042	0.038	0.037
X.8.2027	0.003	0.007	0.021	0.02	0.014	0.020	0.019	0.024	0.033
X40.092	0.872	0.814	0.750	0.707	0.739	0.712	0.631	0.690	0.708
X81.8828	0.032	0.036	0.022	0.028	0.021	0.035	0.034	0.026	0.026

References

[1] D. Cohn, L. Atlas, R. Ladner, Improving generalization with active learning, Mach. Learn. 15 (2) (1994) 201–221.
[2] B. Settles, From theories to queries: active learning in practice, Journal of Machine Learning Research, Workshop on Active Learning and Experimental Design, Workshop and Conference Proceedings 16 (2011) 1–18.
[3] B. Settles, Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, San Rafael, 2012.
[4] H. Su, S. Yin, Z. Huh, T. Kanade, J. Zhu, Interactive cell segmentation based on active and semi-supervised learning, IEEE Trans. Med. Imaging 35 (3) (2016) 762–777.
[5] M. Wang, X. Hua, Active learning in multimedia annotation and retrieval: a survey, ACM Trans. Intell. Syst. Technol. 2 (2) (2011) 10:1–10:21.
[6] T. Wang, Y. Li, H. Xiong, A novel locally active learning method for sar image classification, in: Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium 2014, 2014, pp. 4596–4599.
[7] R. Sousa, A. andPrudêncio, T. Ludermir, C. Soares, Active learning and data manipulation techniques for generating training examples in meta-learning, Neurocomputing 194 (2016) 45–55.
[8] M. Zhang, Y.a. Er, Sequential active learning using meta-cognitive extreme learning machine, Neurocomputing 173 (2016) 835–844.
[9] S. Hao, J. Jiang, Y. Guo, H. Li, Active learning based intervertebral disk classification combining shape and texture similarities, Neurocomputing 101 (2013) 252–257.
[10] D. Pereira-Santos, R. Prudêncio, A. de Carvalho, Empirical investigation of active learning strategies, Neurocomputing (2017). Available online 14 September 2017. In Press, Corrected Proof.
[11] D. Cohn, Neural network exploration using optimal experiment design, Neural Netw. 9 (6) (1996) 1071–1083.
[12] K. Yu, J. Bi, V. Tresp, Active learning via transductive experimental design, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1714–1723.
[13] W. Fu, S. Hao, M. Wang, Active learning on anchorgraph with an improved transductive experimental design, Neurocomputing 2016 (171) (2016) 452–462.
[14] O.M. Aodha, N. Campbell, J. Kautz, G. Brostow, Hierarchical subquery evaluation for active learning on a graph, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), 2014, pp. 564–571.
[15] M. Wang, B. Ni, X. Hua, T. Chua, Assistive tagging: a survey of multimedia tagging with human-computer joint exploration, ACM Comput. Surv. 44 (4) (2012) 25:1–25:24.
[16] X. Yang, Y. Chen, H. Yu, Y. Zhang, Less annotation on active learning using confidence-weighted predictions, Neurocomputing 275 (2018) 1629–1636.
[17] S. Jiang, G. Malkomes, G. Converse, A. Shofner, A. B. Moseley, R. Garnett, Efficient nonmyopic active search, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1714–1723.
[18] J. Wang, P.E., Active learning for penalized logistic regression via sequential experimental design, Neurocomputing 222 (2017) 183–190.
[19] A. Vlachos, A stopping criterion for active learning, Comput. Speech Lang. 22 (3) (2007) 295–312.
[20] J. Zhu, H. Wang, E. Hovy, Multi-criteria-based strategy to stop active learning for data annotation, in: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 2008, pp. 1129–1136.
[21] B. Settles, Active learning literature survey, in: Computer Sciences Technical Report 1648, 2009. University of Wisconsin–Madison.
[22] D.J. Hsu, Algorithms for active learning, 2010 Ph.D. thesis. Columbia University.
[23] M. Culver, D. Kun, S. Scott, Active learning to maximize area under the ROC curve, in: Proceedings of the Sixth International Conference on Data Mining, 2006. ICDM '06., 2006, pp. 149–158.
[24] T.L. Lai, C.Z. Wei, Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems, Ann. Stat. 10 (1) (1982) 154–166.
[25] A.R. Webb, K.D. Copsey, Statistical Pattern Recognition, 3rd Edition, Wiley, 2011.
[26] Z. Wang, Y.C.I. Chang, Sequential estimate for linear regression models with uncertain number of effective variables, Metrika 76 (7) (2013) 949–978.
[27] X.W. Deng, V.R. Joseph, A. Sudjianto, C.F.J. Wu, Active learning through sequential design, with applications to detection of money laundering, J. Am. Stat. Assoc. 104 (487) (2009) 969–981.
[28] R. Azziz, et al., The androgen excess and pcos society criteria for the polycystic ovary syndrome: the complete task force report, Fertil. Steril. 91 (2) (2009) 456–488.
[29] D. Heck, J. Knapp, J. Capdevielle, T. Schatz, G. Thouw, Corsika: a monte carlo code to simulate extesive air showers fzka 6019, Tech. Rep., Forschungszentrum Karlsruhe GmbH(1998). Karlsruhe.
[30] M. Lichman, 2013, UCI machine learning repository.
[31] C.F.J. Wu, Efficient sequential designs with binary data, J. Am. Stat. Assoc. 80 (392) (1985) 974–984.

Yuan-chin Ivan Chang received his Ph.D. degree in Statistics from University of Illinois, Urbana Champaign, USA, and is now a research fellow of Institute of Statistical Science, Academia Sinica and professor of Department of Statistics, National Cheng-Chi University, Taipei, Taiwan.

Ray-Bing Chen is a Professor in the Department of Statistics, National University of Kaohsiung. He received his Ph.D. in Statistics from the University of California, Los Angeles. His research interests include statistical and machine learning, statistical modeling, computer experiment, and optimal design.