Scientific Research Publishing

Open Journal of Statistics, 2022, 12, 711-732
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Model-Free Feature Screening Based on Gini Impurity for Ultrahigh-Dimensional Multiclass Classification

Zhongzheng Wang[1], Guangming Deng[1,2*]

[1]College of Science, Guilin University of Technology, Guilin, China
[2]Applied Statistics Institute, Guilin University of Technology, Guilin, China
Email: *dgm@glut.edu.cn

How to cite this paper: Wang, Z.Z. and Deng, G.M. (2022) Model-Free Feature Screening Based on Gini Impurity for Ultrahigh-Dimensional Multiclass Classification. *Open Journal of Statistics*, **12**, 711-732.
https://doi.org/10.4236/ojs.2022.125042

Received: September 7, 2022
Accepted: October 24, 2022
Published: October 27, 2022

Abstract

It is quite common that both categorical and continuous covariates appear in the data. But, most feature screening methods for ultrahigh-dimensional classification assume the covariates are continuous. And applicable feature screening method is very limited; to handle this non-trivial situation, we propose a model-free feature screening for ultrahigh-dimensional multi-classification with both categorical and continuous covariates. The proposed feature screening method will be based on Gini impurity to evaluate the prediction power of covariates. Under certain regularity conditions, it is proved that the proposed screening procedure possesses the sure screening property and ranking consistency properties. We demonstrate the finite sample performance of the proposed procedure by simulation studies and illustrate using real data analysis.

Keywords

Ultrahigh-Dimensional, Feature Screening, Model-Free, Gini Impurity, Multiclass Classification

1. Introduction

Ultrahigh-dimensional data are commonly available in a wide range of scientific research and applications. Feature screening plays an essential role in the ultrahigh-dimensional data, where Fan and Lv [1] first proposed the sure independence screening (SIS) in the seminal paper. For linear regressions, they showed that the approach based on Pearson correlation learning possesses a sure screening property. That is, even if the number of predictors p can grow much

faster than the number of observations n with $\log p = O\left(n^{\alpha}\right)$ for some $\alpha \in \left(0, \frac{1}{2}\right)$, all relevant predictors can be selected with probability tending to one [2].

Lots of feature screening is the Model-based and Model-free approaches have been developed in recent years, see, for example, Wang [3] proposed forward regression for ultrahigh-dimensional data. Fan and Song [4] applied the maximum marginal likelihood estimates or the maximum marginal likelihood to ultrahigh-dimensional screening in generalized linear model. Fan *et al.* [5] further extend the correlation learning to marginal nonparametric learning. Zhu *et al.* [6] proposed a model-free feature screening approach for ultrahigh-dimensional data. Li *et al.* [7] proposed a robust rank correlation screening method to deal with ultrahigh-dimensional data based on the Kendall τ correlation coefficient. Li *et al.* [8] applied the distance correlation to sure independence screening procedure. He *et al.* [9] proposed a quantile-adaptive framework for nonlinear variable screening with high-dimensional heterogeneous data. Fan *et al.* [10] proposed nonparametric independence screening selects variables by ranking a measure of the nonparametric marginal contributions of each covariate given the exposure variable. Liu *et al.* [11] proposed a feature screening procedure for varying coefficient model based on conditional correlation coefficient. Nandy *et al.* [12] proposed a covariate information number sure independence screening, which used a marginal utility connected to the notion of the traditional Fisher information. Pouyap *et al.* [13] proposed a merge of the features selection methods in order to define the most relevant features in the texture of the vibration signal images.

To address the ultrahigh-dimensional feature screening in classification problem, Fan and Fan [14] proposed the t-test statistic for two-sample mean problem as a marginal utility for feature screening and establish its theoretical properties. Mai and Zou [15] applied the Kolmogorov filter to ultrahigh-dimensional binary classification. Cui *et al.* [16] proposed a screening procedure via used empirical conditional distribution functions. Lai *et al.* [17] proposed a feature screening procedure based on the expected conditional Kolmogorov filter for binary classification problem. However, the above-proposed screening methods assume that the types of data are continuous. For categorical covariates, Huang *et al.* [18] constructed a model-free discrete feature screening method based on the Pearson Chi-square statistics and showed its sure screening property fulfilling (Fan *et al.* [2]). When all the covariates are binary, Ni and Fang [19] proposed a model-free feature screening procedure based on information entropy theory for multi-class classification. Ni *et al.* [20] further proposed a feature screening procedure based on weighting Adjusted Pearson Chi-square for multi-class classification. Sheng and Wang [21] proposed a new model-free feature screening method based on classification accuracy of marginal classifiers for ultrahigh-dimensional classification. Anzarmou *et al.* [22] proposed a new model-free interaction screening method, termed Kendall Interaction Filter (KIF), for the classification in high-

dimensional settings.

Based on the above study of classification models, in this paper, we propose a model-free feature screening for ultrahigh-dimensional multi-classification with both categorical and continuous covariates. The proposed feature screening method will be based on Gini impurity to evaluate the prediction power of covariates. Gini impurity is a non-purity attribute splitting index, which was proposed by Breiman *et al.* [23] and has been widely used in decision tree algorithms such as CART and SPRINT. With regard to categorical covariate screening, we can apply the index of purity gain, which is the same as information gain [19]. Similar to Ni and Fang [19], continuous covariates can be sliced via standard normal quantile. The proposed feature screening procedure is based on purity gain, which is referred to Purity Gain sure independence screening (PG-SIS). Theoretically, the PG-SIS is rigorously proven to enjoy. Fan and Lv [1] proposed sure screening property that ensures all important features can be obtained. Practically, as shown by the simulation results, compared with the existing feature screening method, PG-SIS satisfies the sure screening property.

This paper is organized as follows. Section 2 describes the proposed PG-SIS method in detail. Section 3 establishes its sure screening property. In Section 4, numerical simulations and an example for real data analysis are given to assess sure screening property of our method. Some concluding remarks are given in Section 5 and all the proofs are given in the Appendix.

2. Feature Screening Procedure

We first introduce Gini impurity and purity gain, and then propose the screening procedure based on purity gain.

2.1. Gini Index and Purity Gain

Suppose that Y is a categorical response with R classes $\{1,\cdots,R\}$, and covariate $X=\left(X_1,X_2,\cdots,X_p\right)$ is a vector of p dimension, where each of these components X_k with J_k categories. Where $J_k=\{1,\cdots,J\}$. To introduce the Gini impurity and purity gain, assuming that $Y\in\{1,\cdots,R\}$ and $X_k\in\{1,2,\cdots,J_k\}$. Define $p_r=P(Y=r)$ represents the probability function of a response variable, $w_{k,j}=P(X_k=j)$ represents the probability function of covariates, $p_{k,jr}=P(Y=r\mid X_k=j)$ represents the probability function of response variables under the condition of covariates, where $r=1,2,\cdots,R;\ j=1,2,\cdots,J_k$ and $k=1,2,\cdots,p$. Let $0\times\log 0=0$. Marginal Gini impurity of Y and X respectively is defined as

$$Gini(Y)=1-\sum_{r=1}^{R}p_r^2 \tag{1}$$

$$Gini(X_k)=1-\sum_{j=1}^{J_k}w_{k,j}^2 \tag{2}$$

Conditional Gini impurity is defined as

$$Gini\left(Y \mid X_k\right) = \sum_{j=1}^{J_k} w_{k,j}\left(1-\sum_{r=1}^{R} p_{k,jr}^2\right) \tag{3}$$

Similar to the information gain, the purity gain is defined as

$$\begin{aligned} PG\left(Y \mid X_k\right) &= Gini\left(Y\right) - Gini\left(Y \mid X_k\right) \\ &= 1-\sum_{r=1}^{R} p_r^2 - \sum_{j=1}^{J_k} w_{k,j}\left(1-\sum_{r=1}^{R} p_{k,jr}^2\right) \end{aligned} \tag{4}$$

In the Equation (1), $Gini\left(Y\right)$ is non-negative and acquires its maximum $1-\frac{1}{R}$ if and only if $p_1 = \cdots = p_R = \frac{1}{R}$ by Jensen's inequality [24]. And the $Gini\left(Y \mid X_k\right)$ in Equation (2) is the conditional Gini impurity of Y given $X_k = j$. Further support can be given by the following proposition.

Proposition 2.1. When X_k is a categorical covariable, we can get $PG\left(Y \mid X_k\right) \geq 0$, and X_k and Y are independent if and only if $PG\left(Y \mid X_k\right) = 0$.

For continuous X_k, the conditional Gini impurity can't directly calculate, and purity gain by slicing X into several categories. For a fixed integer $J \geq 2$, let $q_{(j)}$ be the j/J^{th} percentile of X, $j = 1, \cdots, J-1$, $q_{(0)} = -\infty$ and $q_{(J)} = +\infty$. Replacing $w_{k,j}$ and $p_{k,jr}$ in Equation (3) respectively by $w_{k,j} = P\left(X_k \in \left(q_{(j-1)}, q_{(j)}\right]\right)$ and $p_{k,jr} = P\left(Y = r \mid X_k \in \left(q_{(j-1)}, q_{(j)}\right]\right)$, we define conditional Gini impurity based on continuous covariates:

$$\begin{aligned} Gini_J\left(Y \mid X_k\right) &= \sum_{j=1}^{J_k} w_{k,j}\left(1-\sum_{r=1}^{R} p_{k,jr}^2\right) \\ &= \sum_{j=1}^{J_k} P\left(X_k \in \left(q_{(j-1)}, q_{(j)}\right]\right)\left(1-\sum_{r=1}^{R} P\left(Y = r \mid X_k \in \left(q_{(j-1)}, q_{(j)}\right]\right)^2\right) \end{aligned} \tag{5}$$

$$\begin{aligned} PG\left(Y \mid X_k\right) &= Gini\left(Y\right) - Gini\left(Y \mid X_k\right) \\ &= 1-\sum_{r=1}^{R} p_r^2 - \sum_{j=1}^{J_k} w_{k,j}\left(1-\sum_{r=1}^{R} p_{k,jr}^2\right) \\ &= 1-\sum_{r=1}^{R} p_r^2 - \sum_{j=1}^{J_k} P\left(X_k \in \left(q_{(j-1)}, q_{(j)}\right]\right) \\ &\quad \times\left(1-\sum_{r=1}^{R} P\left(Y = r \mid X_k \in \left(q_{(j-1)}, q_{(j)}\right]\right)^2\right) \end{aligned} \tag{6}$$

Proposition 2.2. When X_k is a continuous covariable, we can get $PG_J\left(Y \mid X_k\right) \geq 0$, and X_k and Y are independent if and only if $PG_J\left(Y \mid X_k\right) = 0$.

2.2. Feature Screening Procedure Based on Purity Gain

First, we select a medium scale of simplified model which can almost fully contain D, where $D = \left\{k : F\left(Y \mid x\right)\right.$ functionally depends on X_k for some $\left.Y = r\right\}$, we use an adjusted purity gain index for each pair $\left(Y, X_k\right)$ is as follows:

$$e_k = \frac{1-\sum_{r=1}^{R} p_r^2 - \sum_{j=1}^{J_k} w_{k,j}\left(1-\sum_{r=1}^{R} p_{k,jr}^2\right)}{\log J_k} \tag{7}$$

where $p_r = P(Y=r)$, $w_{k,j} = P(X_k = j)$ and $p_{k,j_r} = P(Y=r \mid X_k = j)$ when X_k is categorical, J_k represents the number of categories of X_k. When X_k is defined as a continuous covariates, J_k represents the number of slices applied to X_k, $w_{k,j} = 1/J_k$, $p_{k,jr} = P\left(Y=r \mid X_k \in \left(q_{k,(j-1)}, q_{k,(j)}\right]\right)$ and $q_{k,(j)}$ represents j/J^{th} percentile of X_k.

There may be more categories of covariates associated with larger purity gain in the original definition of Equation (4), regardless of whether the covariates are important, especially when the number of categories involved in each covariate is different. So Ni and Fang [19] used $\log J_k$ to construct the information gain ratio to solve this problem, where each category of X_k is the same. Similarly, when each category of X_k is the same, for Equation (7), we apply the $\log J_k$ to build an adjusted purity gain index to address the problem, which is also applied to continuous X_k. However, each category of X_k is different, $1-\sum_{j=1}^{J_k} w_{k,j}^2$ is defined as an adjustment factor, which is motivated by the split X_k into several categories via the Decision Tree algorithm.

For sample data $\{x_{i1}, \cdots, x_{ip}, y_i\}$, $i = 1, \cdots, n$, e_k can be easily estimated by

$$\hat{e}_k = \frac{\left(1-\sum_{r=1}^{R} \hat{p}_r^2\right) - \sum_{j=1}^{J_k} \hat{w}_{k,j}\left(1-\sum_{r=1}^{R} \hat{p}_{k,jr}^2\right)}{\log J_k} \tag{8}$$

When X_k is categorical, $\hat{w}_{k,j} = \frac{1}{n}\sum_{i=1}^{n} I\{x_{ik} = j\}$ and $\hat{p}_{k,jr} = \frac{\sum_{i=1}^{n} I\{y_i = r, x_{ik} = j\}}{\sum_{i=1}^{n} I\{x_{ik} = j\}}$.

When X_k is continuous,

$$\hat{w}_{k,j} = \frac{1}{n}\sum_{i=1}^{n} I\left\{x_{ik} \in \left(\hat{q}_{k,(j-1)}, \hat{q}_{k,(j)}\right]\right\}$$

and

$$\hat{p}_{k,j_r} = \frac{\sum_{i=1}^{n} I\left\{y_i = r, x_{ik} \in \left(\hat{q}_{k,(j-1)}, \hat{q}_{k,(j)}\right]\right\}}{\sum_{i=1}^{n} I\left\{x_{ik} \in \left(\hat{q}_{k,(j-1)}, \hat{q}_{k,(j)}\right]\right\}}$$

where $\hat{q}_{k,(j)}$ is the j/J^{th} sample normal percentile of $\{x_{ik}, \cdots, x_{nk}\}$. In either case, $\hat{p}_r = \frac{1}{n}\sum_{i=1}^{n} I\{y_i = r\}$.

We suggest selecting a sub-model: $\hat{D} = \{k : \hat{e}_k \geq cn^{-\tau}, 1 \leq k \leq p\}$. Where both c and τ are predetermined thresholds established via Condition (C2) in Section 3. In practice, we can choose a model:

$$\hat{D} = \{k : \hat{e}_k \text{ is among the top of } d \text{ largest of all}\}$$

where $d = [n/\log n]$.

3. Feature Screening Property

In this section, we establish the sure screening property of PG-SIS. Based on Ni

and Fang [19] proposed sure independence screening theories, the following conditions are assumed.

Condition 1 (C1). There exist two positive constants c_1 and c_2 such that, $c_1/R \le p_r \le c_2/R$, $c_1 + c_2 \le R$, $c_1/R \le p_{k,jr} \le c_2/R$ and $c_1/J_k \le w_{k,j} \le c_2/J_k$ for every $1 \le j \le J_k$, $1 \le r \le R$ and $1 \le k \le p$.

Condition 2 (C2). There exist a positive constant $c > 0$ and $0 \le \tau < 1/2$ such that $\min_{k \in D} e_k \ge 2cn^{-\tau}$.

Condition 3 (C3). $R = O\left(n^{\varepsilon}\right)$, $J = \max_{1 \le k \le P} J_k = O\left(n^{k}\right)$, where $\varepsilon \ge 0$, $\kappa \ge 0$ and $2\tau + 2\varepsilon + 2\kappa < 1$.

Condition 4 (C4). There exist a positive constant c_3, such that $0 < f_k\left(x \mid Y = r\right) < c_3$ for any $1 \le r \le R$, and x in the domain of X_k, where $f_k\left(x \mid Y = r\right)$ is the Lebesgue density function of X_k conditional on $Y = r$.

Condition 5 (C5). There exist a positive constant c_4 and $0 \le \rho < 1/2$ such that $f_k\left(x\right) \ge c_4 n^{-\rho}$ for any $1 \le k \le p$ and x in the domain of X_k, where $f_k\left(x\right)$ is the Lebesgue density function of X_k. Furthermore, $f_k\left(x\right)$ is continuous in the domain of X_k.

Condition 6 (C6). $R = O\left(n^{\varepsilon}\right)$, $J = \max_{1 \le k \le P} J_k = O\left(n^{\kappa}\right)$, where $2\tau + 2\varepsilon + 2\kappa + 2\rho < 1$ and $\varepsilon \ge 0, \kappa \ge 0$.

Condition 7 (C7). $\liminf_{p \to \infty} \left\{\min_{k \in D} e_k - \max_{k \in I} e_k\right\} \ge \delta$, where $\delta > 0$ is a constant.

Condition (C1) guarantees that the proportion of each class of variables cannot be either extremely small or extremely large. Similar assumption is also made in condition (C1) in Huang *et al.* [18] and Cui *et al.* [16]. According to Fan and Lv [1] and Cui *et al.* [16], Condition (C2) allows the minimum true signal to disappear to zero in the order of $n^{-\tau}$ as the sample size goes to infinity. According to [19] Condition (C3) provides the covariates to diverge with a certain order and the number of classes for the response, and Condition (C6) modifies Condition (C3) a little bit. To ensures the sample percentiles are close to the true percentiles, Condition (C4) rules out the extreme case that some X_k put heavy mass in a small range. Condition (C5) asks for the $n^{-\rho}$ as lower bound to the density. According to [16] and Zhu *et al.* [6] proposed ranking consistency property, we need to assume the inactive covariate subset $I = \{1, \cdots, p\} \setminus D$, then Condition (C7) is established.

Theorem 3.1. (Sure screening property) Under conditions (C1) to (C3), if all the covariates are categorical, we get:

$$P\left(D \subseteq \hat{D}\right) \ge 1 - O\left(p \exp\left\{-bn^{1-(2\tau+2\varepsilon+2\kappa)} + (\varepsilon + \kappa)\log n\right\}\right)$$

Theorem 3.2. (Sure screening property) Under conditions (C4) to (C6), when the covariates are composed of continuous and categorical variables, we get:

$$P\left(D \subseteq \hat{D}\right) \ge 1 - O\left(p \exp\left\{-bn^{1-(2\tau+2\varepsilon+2\kappa+2\rho)} + (\varepsilon + \kappa)\log n\right\}\right)$$

where b is a positive constant. If $\log p = O\left(n^{\alpha}\right)$ and $\alpha < 1 - \left(2\tau + 2\varepsilon + 2\kappa + 2\rho\right)$,

PG-SIS has a sure screening property.

Theorem 3.3. (Ranking consistency property) Under conditions (C1), (C4), (C5) and (C7), if $\log\frac{RJ}{\log n}=O(1)$ and $\frac{\max\{\log P,\log n\}R^4J^4}{n^{1-2\rho}}=o(1)$, then $\liminf_{n\to\infty}\{\min_{k\in D}\hat{e}_k-\max_{k\in I}\hat{e}_k\}>0$, *a.s.*

Theorem 3.3 testifies that the proposed screening index can separate active and inactive covariates well in the sample level.

4. Numerical Studies

4.1. Simulation Results

In this subsection, we carry out three simulation studies to demonstrate the finite sample performance of our group screen methods described in Section 2. We compare PG-SIS with IG-SIS [19] and APC-SIS in performance via the below evaluation criteria: MMS, minimal model size, consists of all active covariates, the results generally existing 5%, 25%, 50%, 75%, 95% of MMS; CP1, CP2 and CP3 respectively represent the probability that the given model size $[n/\log n]$, $2[n/\log n]$ and $3[n/\log n]$ cover all active covariates, while CPa indicates whether the indicators of the selected model cover all active covariates.

Model 1: categorical covariates and binary response

We first consider the response variables of different categories. According to [19], we assume a model which response y_i is binary in which $R=2$, and all the covariates are categorical. We think about two distributions for y_i:

1) Balanced, $P(y_i=r)=1/2$;

2) Unbalanced, $p_r=2\left[1+\frac{R-r}{R-1}\right]\Big/3R$ with $\max_{1\le r\le R}p_r=2\min_{1\le r\le R}p_r$.

The true model is defined at $D=\{1,2,\cdots,20\}$ with $d_0=|D|=20$. Condition on y_i, latent variable is generated as $z_i=(z_{i,1},\cdots,z_{i,p})$, where $z_{i,k}$ $N(\mu_{rk},1)$, $1\le k\le p$. Then, we construct active covariates:

1) If $k>d_0$, then $\mu_{rk}=0$;

2) If $k\le d_0$ and $r=1$, then $\mu_{rk}=-0.5$;

3) If $k\le d_0$ and $r=2$, then $\mu_{rk}=0.5$.

Next, we apply the quantile of the standard normal distribution to generate covariates. The specific approach is as follows:

1) When k as odd number, that is $x_{i,k}=I\left(z_{i,k}>z_{\left(\frac{j}{2}\right)}\right)+1$;

2) When k as even number, that is $x_{i,k}=I\left(z_{i,k}>z_{\left(\frac{j}{5}\right)}\right)+1$;

Where αth percentile of the standard normal distribution is $z_{(\alpha)}$.

Thus, amongst all p covariates, the covariates of two categories and five categories account for half, respectively. Similar to [20], we consider $p=1000,5000$ and $n=200,400$ in this model.

Table 1 reports the evaluation criteria over 100 simulations for Model 1. We

Table 1. Simulation results for example 1.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
Balanced Y, $n = 200$, $p = 1000$									
PG-SIS	20.0	20.0	21.0	22.0	23.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	21.0	22.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 400$, $p = 1000$									
PG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 200$, $p = 5000$									
PG-SIS	22.0	24.0	25.5	27.3	32.1	0.982	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.8	21.0	23.0	28.0	0.996	1.000	0.000	0.000
Balanced Y, $n = 400$, $p = 5000$									
PG-SIS	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000
Unbalanced Y, $n = 200$, $p = 1000$									
PG-SIS	20.0	20.0	21.0	22.0	23.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
APC-SIS	21.0	23.0	25.0	27.0	30.1	0.984	1.000	0.000	0.000
Unbalanced Y, $n = 400$, $p = 1000$									
PG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000
Unbalanced Y, $n = 200$, $p = 5000$									
PG-SIS	21.0	24.0	26.0	28.0	31.0	0.975	1.000	0.995	1.000
IG-SIS	20.0	20.0	20.5	21.0	23.1	1.000	1.000	1.000	1.000
APC-SIS	29.0	35.0	43.0	53.0	89.1	0.914	0.983	0.000	0.000
Unbalanced Y, $n = 400$, $p = 5000$									
PG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000

can see the following: The results argue that the proposed PG-SIS works quite well. When the sample size n increases, PG-SIS is close to the true model size $d_0 = 20$ in MMS, and both increase to 1 in coverage probability. MMS in unbalanced response is better than in balanced response in performance via comparing the response of different structures. The performances of PG-SIS and IG-SIS are quite close, and PG-SIS is slightly better than APC-SIS in higher coverage probabilities.

Model 2: categorical covariates and multi-class response

We consider more covariate classification, and response y_i is multi-class which $R = 10$. We think about y_i of two distributions:

1) Balanced, $P(y_i = r) = 1/R$;

2) Unbalanced, $p_r = 2\left[1 + \dfrac{R-r}{R-1}\right] \Big/ 3R$ with $\max_{1\le r\le R} p_r = 2\min_{1\le r\le R} p_r$.

Among the $p = 2000$ covariates, the minimum set of active covariate set is $X^D = \{X_{200}, X_{400}, X_{600}, X_{800}, X_{1000}, \cdots, X_{2000}\}$ with the number of active covariates $d_0 = 10$. Condition on y_i, latent variable is generated as $z_i = (z_{i,1}, \cdots, z_{i,p})$, for covariates X_k, $x_{i,k} = f_k(\varepsilon_{i,k} + \mu_{i,k})$, where $\varepsilon_{i,k} \sim N(0,1)$ and $f_k(\cdot)$ represents a quantile function of standard normal distribution. Then, we construct active covariates via defining $\mu_{i,k}$:

1) If $X \in X^D$ and $y_i = r$, then $\mu_{i,k} = 1.5 \times (-0.9)^r$;

2) If $X \notin X^D$, then $\mu_{i,k} = 0$;

Next, we apply the $f_k(\cdot)$ to generate covariates, and take $p = 2000$, $n = 300, 400, 500$ in this model. The specific approach is as follows:

1) For $1 \le k \le 400$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\left(\frac{j}{2}\right)}\right) + 1$;

2) For $400 < k \le 800$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\left(\frac{j}{4}\right)}\right) + 1$;

3) For $800 < k \le 1200$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\left(\frac{j}{6}\right)}\right) + 1$;

4) For $1200 < k \le 1600$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\left(\frac{j}{8}\right)}\right) + 1$;

5) For $1600 < k \le 2000$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\left(\frac{j}{10}\right)}\right) + 1$;

Thus, amongst all the p covariates, the covariates of two categories, four categories, six categories, eight categories and ten categories account for one-fifth each.

Table 2 reports the evaluation criteria over 100 simulations for Model 2. We can see the following: Two methods in performance under Model 1 is worse than Model 2. When the model is more intricate, PG-SIS in performance is close to IG-SIS. Particularly, PG-SIS and IG-SIS have a slightly small MMS under a small sample size n. When the sample size n increases, PG-SIS is close to

Table 2. Simulation results for example 2.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
Balanced Y, $n = 300$, $p = 2000$									
PG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	11.0	12.0	14.0	20.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 400$, $p = 2000$									
PG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 500$, $p = 2000$									
PG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000
Unbalanced Y, $n = 300$, $p = 2000$									
PG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	11.0	1.000	1.000	0.000	0.000
Unbalanced Y, $n = 400$, $p = 2000$									
PG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000
Unbalanced Y, $n = 500$, $p = 2000$									
PG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000

$d_0 = 10$ in MMS, and both increase to 1 in coverage probability. Four indexes of coverage probability of APC-SIS are worse than that of PG-SIS when the sample $n = 200$. MMS in unbalanced response is better than in balanced response in performance via comparing the response of different structures. Furthermore, PG-SIS and IG-SIS are more robust in performance because the fluctuation range in MMS is small.

Model 3: continuous and categorical covariates

At last, among the covariates that are both continuous and categorical, we as-

sume a more complex example, where response y_i is multi-class which $R=4$. We think about y_i of two distributions:

1) Balanced, $p_r = P(y_i = r) = 1/R$;

2) Unbalanced, $p_r = 2\left[1+\frac{R-r}{R-1}\right]\Big/3R$ with $\max_{1\le r\le R} p_r = 2\min_{1\le r\le R} p_r$.

In this model, we take $p=5000$, $n=400,600,800$. The true model is defined at $X^D = \left\{X_k : k = \left[\frac{k'p}{20}\right], k' = 1,\cdots,20\right\}$ with $d_0 = 20$. Condition on y_i, latent variable is generated as $z_i = (z_{i,1},\cdots,z_{i,p})$. For covariates X_k, $z_{i,k}$ $N(\mu_{i,k},1), 1\le k\le p$, where $u_{ik} = (-1)^r \theta_{rk}$ when $y_i = r$ and $k\in D$. According to Ni and Fang [19], θ_{rk} is given in Table 3. $\mu_{i,k}=0$ when $k\notin D$. To generate X_k:

For $k \le \left[\frac{5p}{20}\right]$, then $x_{i,k} = j$, if $z_{i,k} \in \left(z_{(j-1)/4}, z_{j/4}\right], j=1,2,3,4$;

For $\left[\frac{5p}{20}\right] < k \le \left[\frac{10p}{20}\right]$, then $x_{i,k} = j$, if $z_{ik} \in \left(z_{(j-1)/10}, z_{j/10}\right], j=1,\cdots,10$;

For $\left[\frac{10p}{20}\right] < k \le p$, then $x_{i,k} = z_{i,k}$.

Thus, amongst all the p covariates, the covariates of four categories and ten categories account for one-fifth, respectively, the other covariates are continuous. Similarly, there respectively are 5 in four categories and ten categories in the active covariates, and the rest of active covariates are continuous accounting for half. For continuous covariates, we applied different slices $J_k = 4,8,10$. The corresponding approaches are defined as PG-SIS-4, IG-SIS-4, APC-SIS = 4, PG-SIS-8, IG-SIS-8, APC-SIS-8, PG-SIS-10, IG-SIS-10 and APC-SIS-10. Table 4 and Table 5 show the simulation results with over 100 simulations for balanced and unbalanced case, respectively. We can see the following: When the sample size n increases, PG-SIS is close to $d_0 = 20$ in MMS, and both increase to 1 in coverage probability. And coverage probability of PG-SIS is close to IG-SIS in five indexes. Therefore, it is proved that the PG-SIS has the characteristics of feature screening. MMS in unbalanced response is better than in balanced response in performance via comparing the response of different structures. Furthermore, PG-SIS and IG-SIS are robust in performance because the fluctuation range in MMS is small for two types of responses. When different slices are

Table 3. Parameter specification of Model 3.

θ_{rk}	K									
	1	2	3	4	5	6	7	8	9	10
$r=1$	0.2	0.8	0.7	0.2	0.2	0.9	0.1	0.1	0.7	0.7
$r=2$	0.9	0.3	0.3	0.7	0.8	0.4	0.7	0.6	0.4	0.4
$r=3$	0.1	0.9	0.9	0.1	0.3	0.1	0.4	0.3	0.6	0.6
$r=4$	0.7	0.2	0.2	0.6	0.7	0.6	0.8	0.9	0.1	0.1

Table 4. Simulation results for example 3: balanced Y.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	Balanced Y, $n = 400$, $p = 5000$								
PG-SIS-4	27.0	34.0	38.0	47.0	68.1	0.984	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	21.0	21.1	1.000	1.000	1.000	1.000
PG-SIS-8	22.0	24.0	27.0	29.3	35.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	21.0	21.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	21.0	22.0	1.000	1.000	1.000	1.000
PG-SIS-10	22.0	24.0	27.0	29.0	34.1	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	21.0	22.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	21.0	21.1	1.000	1.000	1.000	1.000
	Balanced Y, $n = 600$, $p = 5000$								
PG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
	Balanced Y, $n = 800$, $p = 5000$								
PG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000

Table 5. Simulation results for example 3: unbalanced Y.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	Unbalanced Y, $n = 400$, $p = 5000$								
PG-SIS-4	39.2	40.0	41.0	42.0	42.8	1.000	1.000	1.000	1.000
IG-SIS-4	20.1	20.5	21.0	21.5	21.9	1.000	1.000	1.000	1.000
APC-SIS-4	20.2	20.8	21.5	22.3	22.9	1.000	1.000	1.000	1.000
PG-SIS-8	25.2	26.0	27.0	28.0	28.8	1.000	1.000	1.000	1.000
IG-SIS-8	20.2	20.8	21.5	22.3	22.9	1.000	1.000	1.000	1.000
APC-SIS-8	20.2	21.0	22.0	23.0	23.8	1.000	1.000	1.000	1.000
PG-SIS-10	25.2	26.0	27.0	28.0	28.8	1.000	1.000	1.000	1.000
IG-SIS-10	20.3	21.3	22.5	23.8	24.8	1.000	1.000	1.000	1.000
APC-SIS-10	20.2	20.8	21.5	22.3	22.9	1.000	1.000	1.000	1.000
	Unbalanced Y, $n = 600$, $p = 5000$								
PG-SIS-4	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
	Unbalanced Y, $n = 800$, $p = 5000$								
PG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
PG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000

Table 6. Analysis results for real data example.

	screening method	response			
		1	2	3	4
classification method		**SVM**			
G-mean (training data)	APC-SIS	1.0000	0.9304	0.9709	0.9713
	IG-SIS	1.0000	0.9025	0.9093	1.0000
	PG-SIS	0.9853	0.9378	0.9514	0.9946
G-mean (test data)	APC-SIS	0.9775	0.9913	0.9564	0.9379
	IG-SIS	1.0000	0.9439	0.8736	0.9922
	PG-SIS	0.9678	0.9779	0.9173	0.9739
F-measure (training data)	APC-SIS	0.7673	0.5958	0.7124	0.7101
	IG-SIS	0.6924	0.3969	0.4121	0.7004
	PG-SIS	0.7307	0.6095	0.6511	0.7469
F-measure (test data)	APC-SIS	0.5424	0.3233	0.5333	0.4033
	IG-SIS	0.5850	0.1667	0.2167	0.5505
	PG-SIS	0.4533	0.2667	0.3967	0.5057
classification method		**DT**			
G-mean (training data)	APC-SIS	0.9909	0.8898	0.9489	0.9872
	IG-SIS	0.9945	0.8743	0.9437	0.9917
	PG-SIS	0.9815	0.8902	0.9578	0.9835
G-mean (test data)	APC-SIS	0.9913	0.9626	0.9371	0.9774
	IG-SIS	0.9913	0.9200	0.9371	0.9862
	PG-SIS	0.9862	0.8963	0.8838	0.9609
F-measure (training data)	APC-SIS	0.6743	0.2915	0.5757	0.6648
	IG-SIS	0.6668	0.1792	0.5466	0.6613
	PG-SIS	0.6424	0.2807	0.5825	0.6468
F-measure (test data)	APC-SIS	0.5457	0.1967	0.4000	0.5367
	IG-SIS	0.5790	0.0500	0.4333	0.5471
	PG-SIS	0.4667	0.0500	0.2933	0.4333
classification method		**RF**			
G-mean (training data)	APC-SIS	1.0000	0.9458	1.0000	1.0000
	IG-SIS	1.0000	0.9458	1.0000	1.0000
	PG-SIS	0.9923	0.9421	0.9782	1.0000

Continued

G-mean (test data)	APC-SIS	1.0000	0.9807	0.9540	0.9835
	IG-SIS	1.0000	0.9894	0.9523	0.9453
	PG-SIS	0.9871	0.9524	0.9384	0.9774
F-measure (training data)	APC-SIS	1.0000	1.0000	1.0000	1.0000
	IG-SIS	1.0000	1.0000	1.0000	1.0000
	PG-SIS	0.8603	0.7671	0.8417	0.8725
F-measure (test data)	APC-SIS	0.6624	0.3500	0.6300	0.6300
	IG-SIS	0.6124	0.3567	0.5733	0.4667
	PG-SIS	0.5967	0.2833	0.4933	0.5733

applied in continuous covariates, PG-SIS and IG-SIS are better in five indexes of coverage probability and MMS in performance via comparing the response of different structures. Therefore, three methods are independent of the number of slices in performance.

4.2. Real Data

In this subsection, we analyse a real data set from the feature selection database of Arizona State University (http://featureselection.asu.edu/). The GLIOMA biological data includes 50 samples and 4434 features, which is unbalanced due to the response variable. Every class is 14, 7, 14, 15, and the covariates are not only continuous, but also multiclass. We randomly divided the data into two parts where 90 percent of the data represent training data and 10 percent of the data represents test data. The sample size of training data and test data respectively are $n = 45$ and $n = 5$. The dimension of both training data and test data are $p = 4434$.

We apply a ten-fold cross-validation to eliminate different training data that cause the model accuracy problems. To PG-SIS, IG-SIS and APC-SIS, we use three classification approaches, which are Support Vector Machine [25], Random Forest (RF) and Decision Tree (DT) [26] to them via the chose active covariates.

In training data, we use the G-mean and F-measure [27] evaluation, the same is true for test data. PG-SIS in performance for unbalanced data is reported in Table 6. In all classification methods, PG-SIS is the best in performance, where G-mean of PG-SIS is more closed to 1 than the other two methods. In a word, the proposed PG-SIS performs better.

5. Conclusions

In the data, there are continuous and categorical covariates, and the response is categorical, which is very common in practice, but the applicable screening methods are very limited. We propose a PG-SIS procedure based on Gini impurity

to effectively screen covariates. PG-SIS has a sure screening property and ranking consistency property theoretically and is model free. When the numbers of categories of all the covariates are the same and different, PG-SIS is quite similar to IG-SIS in performance, which can be shown in simulation.

The features screening reports some difficulties based on missing data. In the future, based on the classification model, we intend to propose a new feature screening to either the missing variable or the response variable can be selected.

Acknowledgements

The work was supported by National Natural Science Foundation of China [grant number 71963008].

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Fan, J.Q. and Lv, J.C. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **70**, 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[2] Fan, J.Q., Samworth, R. and Wu, Y.C. (2009) Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research*, **10**, 2013-2038. http://arxiv.org/abs/0812.3201

[3] Wang, H.S. (2009) Forward Regression for Ultra-High Dimensional Variable Screening. *Journal of the American Statistical Association*, **104**, 1512-1524. https://doi.org/10.1198/jasa.2008.tm08516

[4] Fan, J.Q. and Song, R. (2010) Sure Independence Screening in Generalized Linear Models with NP-Dimensionality. *Annals of Statistics*, **38**, 3567-3604. https://doi.org/10.1214/10-AOS798

[5] Fan, J.Q., Feng, Y. and Song, R. (2011) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models. *Journal of the American Statistical Association*, **106**, 544-557. https://doi.org/10.1198/jasa.2011.tm09779

[6] Zhu, L.P., Li, L.X., Li, R.Z. and Zhu, L.X. (2011) Model-Free Feature Screening for Ultrahigh-Dimensional Data. *Journal of the American Statistical Association*, **106**, 1464-1475. https://doi.org/10.1198/jasa.2011.tm10563

[7] Li, G.R., Peng, H., Zhang, J. and Zhu, L.X. (2012) Robust Rank Correlation Based Screening. *Annals of Statistics*, **40**, 1846-1877. https://doi.org/10.1214/12-AOS1024

[8] Li, R.Z., Zhong, W. and Zhu, L.P. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. https://doi.org/10.1080/01621459.2012.695654

[9] He, X.M., Wang, L. and Hong, H.G. (2013) Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data. *Annals of Statistics*, **41**, 342-369. https://doi.org/10.1214/13-AOS1087

[10] Fan, J.Q., Ma, Y.B. and Dai, W. (2014) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models. *Journal of the American Statistical Association*, **109**, 1270-1284.

https://doi.org/10.1080/01621459.2013.879828

[11] Liu, J.Y., Li, R.Z. and Wu, R.L. (2014) Feature Selection for Varying Coefficient Models with Ultrahigh-Dimensional Covariates. *Statistics & Probability Letters*, **109**, 266-274. https://doi.org/10.1080/01621459.2013.850086

[12] Nandy, D., Chiaromonte, F. and Li, R.Z. (2021) Covariate Information Number for Feature Screening in Ultrahigh-Dimensional Supervised Problems. *Journal of the American Statistical Association*, **117**, 1516-1529. https://doi.org/10.1080/01621459.2020.1864380

[13] Pouyap, M., Bit joka, L., Mfoumou, E. and Toko, D. (2021) Improved Bearing Fault Diagnosis by Feature Extraction Based on GLCM, Fusion of Selection Methods, and Multiclass-Naive Bayes Classification. *Journal of Signal and Information Processing*, **12**, 71-85. https://doi.org/10.4236/jsip.2021.124004

[14] Fan, J.Q. and Fan, Y.Y. (2008) High-Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics*, **36**, 2605-2637. https://doi.org/10.1214/07-AOS504

[15] Mai, Q. and Zou, H. (2013) The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification. *Biometrika*, **100**, 229-234. https://doi.org/10.1093/biomet/ass062

[16] Cui, H.J., Li, R.Z. and Zhong, W. (2015) Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, **110**, 630-641. https://doi.org/10.1080/01621459.2014.920256

[17] Lai, P., Song, F.L., Chen, K.W. and Liu, Z. (2017) Model Free Feature Screening with Dependent Variable in Ultrahigh Dimensional Binary Classification. *Statistics & Probability Letters*, **125**, 141-148. https://doi.org/10.1016/j.spl.2017.02.011

[18] Huang, D.Y., Li, R.Z. and Wang, H.S. (2014) Feature Screening for Ultrahigh Dimensional Categorical Data with Applications. *Journal of Business & Economic Statistics*, **32**, 237-244. https://doi.org/10.1080/07350015.2013.863158

[19] Ni, L. and Fang, F. (2016) Entropy-Based Model-Free Feature Screening for Ultrahigh-Dimensional Multiclass Classification. *Journal of Nonparametric Statistics*, **28**, 515-530. https://doi.org/10.1080/10485252.2016.1167206

[20] Ni, L., Fang, F. and Wan, F.J. (2017) Adjusted Pearson Chi-Square Feature Screening for Multi-Classification with Ultrahigh Dimensional Data. *Metrika*, **80**, 805-828. https://doi.org/10.1007/s00184-017-0629-9

[21] Sheng, Y. and Wang, Q.H. (2020) Model-Free Feature Screening for Ultrahigh Dimensional Classification. *Journal of Multivariate Analysis*, **178**, 1-12. https://doi.org/10.1016/j.jmva.2020.104618

[22] Anzarmou, Y., Mkhadri, A. and Oualkacha, K. (2022) The Kendall Interaction Filter for Variable Interaction Screening in High Dimensional Classification Problems. *Journal of Applied Statistics*, 1-19. https://doi.org/10.1080/02664763.2022.2031125

[23] Breiman, L., Friedman, J.H., Stone, C.J. and Olshen, R.A. (1984) Classification and Regression Trees. Wadsworth International Group, Belmont.

[24] Marco, T. (2012) Lectures on Probability Theory and Mathematical Statistics. CreateSpace Independent Publishing Platform, Scotts Valley.

[25] Suykens, J.A.K. and Vandewalle, J. (1999) Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, **9**, 293-300. https://doi.org/10.1023/A:1018628609742

[26] Lantz, B. (2015) Machine Learning with R. Mathematical & Statistical Software. Packt Publishing, Birmingham.

[27] He, H.J. and Deng, G.M. (2022) Grouped Feature Screening for Ultra-High Dimensional Data for the Classification Model. *Journal of Statistical Computation and Simulation*, 1-24.

Appendix

Proof of Proposition 2.1. To prove Proposition 2.1, we need to define $f(x)=x^2$, proved to be close Ni and Fang [19]. By Jensen's inequality [24],

$$\begin{aligned}\sum_{j=1}^{J_k} w_{k,j}\sum_{r=1}^{R} p_{k,jr}^2 &= \sum_{r=1}^{R}\left[\sum_{j=1}^{J_k} w_{k,j} f\left(p_{k,jr}\right)\right]\\ &\geq \sum_{r=1}^{R} f\left(\sum_{j=1}^{J_k} w_{k,j} p_{k,jr}\right)\\ &= \sum_{r=1}^{R} f\left(\sum_{j=1}^{J_k} P\left(X_k=j\right)P\left(Y=r\mid X_k=j\right)\right)\\ &= \sum_{r=1}^{R} p_r^2\end{aligned}$$

then

$$\begin{aligned}PG &= 1-\sum_{r=1}^{R} p_r^2-\sum_{j=1}^{J_k} w_{k,j}\left(1-\sum_{r=1}^{R} p_{k,jr}^2\right)\\ &= 1-\sum_{r=1}^{R} p_r^2-\sum_{j=1}^{J_k} w_{k,j}+\sum_{j=1}^{J_k} w_{k,j}\sum_{r=1}^{R} p_{k,jr}^2\\ &= \sum_{j=1}^{J_k} w_{k,j}\sum_{r=1}^{R} p_{k,jr}^2-\sum_{r=1}^{R} p_r^2 \geq 0\end{aligned}$$

The above equation holds if and only if $p_{k,jr}=p_{k,j'r}$, for any $1\leq r\leq R$ and $1\leq j\leq j'\leq J$. That is, X_k and Y are independent.

Proof of Proposition 2.2. From the same proof as Proposition 2.1, we can get $PG_J\left(Y\mid X_k\right)\geq 0$ holds if and only if $p_{k,j'r}=p_{k,j'r}$ for any $1\leq r\leq R$ and $1\leq j\leq j'\leq J$, that is $P\left(Y=r\mid X\in\left(q_{(j-1)},q_{(j)}\right]\right)=P\left(Y=r\mid X\in\left(q_{(j'-1)},q_{(j')}\right]\right)$. So when X_k and Y are independent, $PG_J\left(Y\mid X_k\right)=0$. On the other hand, when $PG_J\left(Y\mid X_k\right)=0$ for any J, we need to show that $P\left(X\leq x\mid Y=r\right)$ does not depend on r for any x in the domain of X. Proposition 2.2 has been proven by [19], so the proof is omitted here.

Lemma 1 (Bernstein inequality). If $Z_1,\cdots,Z_n$ is an independent random variable with a mean value of 0 and bounded supporter is $[-M,M]$, then the inequality:

$$P\left(\left|\sum_{i=1}^{n} Z_i\right|>t\right)\leq 2\exp\left\{-\frac{t^2}{2\left(v+\frac{Mt}{3}\right)}\right\}$$

where $v\geq Var\left(\sum_{i=1}^{n} Z_i\right)$

Lemma 2. For discrete covariates X_k and discrete response Y, we have the following three inequalities:

1) $P\left(\left|\hat{p}_r-p_r\right|>t\right)\leq 2\exp\left\{-\frac{6nt^2}{3+4t}\right\}$

2) $P\left(\left|\hat{w}_{k,j}-w_{k,j}\right|>t\right)\leq 2\exp\left\{-\frac{6nt^2}{3+4t}\right\}$

3) $P\left(\left|\hat{p}_{k,jr}-p_{k,jr}\right|>t\right)\le 2\exp\left\{-\frac{6nt^2}{3+4t}\right\}$

Proof of Lemma 2. Three inequalities are similar in the proofs, where inequality (1) and inequality (2) have been given at [27]. The following is the proof of inequality (3).

$$\hat{p}_{k,jr}=\frac{\sum_{i=1}^{n} I\left\{y_i=r, X_{i,k}=j\right\}}{\sum_{i=1}^{n} I\left\{X_{i,k}=j\right\}}$$

The expectation of $\hat{p}_{k,jr}$ is

$$E\left(\hat{p}_{k,jr}\right)=E\left(\frac{\sum_{i=1}^{n} I\left\{y_i=r, X_{i,k}=j\right\}}{\sum_{i=1}^{n} I\left\{X_{i,k}=j\right\}}\right)=E\left(\frac{\sum_{i=1}^{n} y_i=r, X_{i,k}=j}{\sum_{i=1}^{n} X_{i,k}=j}\right)=p_{k,jr}$$

Let $Z_i=I\left\{y_i=r\mid X_{i,k}=j\right\}-p_{k,jr}$,

$Var\left(\sum_{i=1}^{n} Z_i\right)=nVar\left(Z_i\right)=np_{k,jr}\left(1-p_{k,jr}\right)\le\frac{n}{4}$ is known, then

$$P\left(\left|\hat{p}_{k,jr}-p_{k,jr}\right|>t\right)=P\left(\left|n^{-1}\sum_{i=1}^{n} Z_i\right|>t\right)=P\left(\left|\sum_{i=1}^{n} Z_i\right|>nt\right)$$

$$\le 2\exp\left\{-\frac{n^2t^2}{2\left(\frac{n}{4}+\frac{nt}{3}\right)}\right\}\le 2\exp\left\{-\frac{6nt^2}{3+4t}\right\}$$

According to Bernstein inequality, the formula is held.

Lemma 3. With regard to discrete covariates X_k and discrete response Y, for any $0<\varepsilon<1$, under condition (C1), we have

$$P\left(\left|\hat{e}_k-e_k\right|>2\varepsilon\right)\le O\left(RJ^3\right)\exp\left\{-c_5\frac{n\varepsilon^2}{R^2J^6}\right\},$$

where c_5 represents a positive constant.

Proof of Lemma 3. By e_k and $\hat{e}_k$ in Section 2.2, we have

$$\begin{aligned}
&\log J_K\left(\hat{e}_k-e_k\right)\\
&=\left[\left(1-\sum_{r=1}^{R}\hat{p}_r^2\right)-\sum_{j=1}^{J_k}\hat{w}_{k,j}\left(1-\sum_{r=1}^{R}\hat{p}_{k,jr}^2\right)\right]-\left[1-\sum_{r=1}^{R}p_r^2-\sum_{j=1}^{J_k}w_{k,j}\left(1-\sum_{r=1}^{R}p_{k,jr}^2\right)\right]\\
&=\left(\sum_{r=1}^{R}p_r^2-\sum_{r=1}^{R}\hat{p}_r^2\right)+\left(\sum_{j=1}^{J_k}w_{k,j}-\sum_{j=1}^{J_k}\hat{w}_{k,j}\right)+\left(\sum_{j=1}^{J_k}\hat{w}_{k,j}\sum_{r=1}^{R}\hat{p}_{k,jr}^2-\sum_{j=1}^{J_k}w_{k,j}\sum_{r=1}^{R}p_{k,jr}^2\right)\\
&=\sum_{r=1}^{R}\left(p_r^2-\hat{p}_r^2\right)+\sum_{j=1}^{J_k}\left(w_{k,j}-\hat{w}_{k,j}\right)+\sum_{j=1}^{J_k}\sum_{r=1}^{R}\left(\hat{w}_{k,j}\hat{p}_{k,jr}^2-w_{k,j}p_{k,jr}^2\right)\\
&=\sum_{r=1}^{R}\left(p_r-\hat{p}_r\right)\left(p_r+\hat{p}_r\right)+\sum_{j=1}^{J_k}\left(w_{k,j}-\hat{w}_{k,j}\right)\\
&\quad+\sum_{j=1}^{J_k}\sum_{r=1}^{R}\left[\left(\hat{w}_{k,j}\hat{p}_{k,jr}+w_{k,j}p_{k,jr}\right)\left(\hat{p}_{k,jr}-p_{k,jr}\right)+\hat{p}_{k,jr}p_{k,jr}\left(w_{k,j}-\hat{w}_{k,j}\right)\right]\\
&=I_1+I_2+I_3
\end{aligned}$$

Since $\log J\ge\log 2\ge 0.5$, we have

$$P\left(\left|\hat{e}_k - e_k\right| > \varepsilon\right) \le P\left(\left|I_1\right| > \frac{\varepsilon}{3}\right) + P\left(\left|I_2\right| > \frac{\varepsilon}{3}\right) + P\left(\left|I_3\right| > \frac{\varepsilon}{3}\right)$$

For I_1, we have

$$\begin{aligned} P\left(\left|I_1\right| > \frac{\varepsilon}{3}\right) &\le \sum_{r=1}^{R} P\left(\left|\left(p_r - \hat{p}_r\right)\left(p_r + \hat{p}_r\right)\right| > \frac{\varepsilon}{3}\right) \\ &\le \sum_{r=1}^{R} P\left(\left|p_r - \hat{p}_r\right| > \frac{c_1\varepsilon}{3RJ^3}\right) \\ &\le RJ^3 2\exp\left\{-\frac{6n\left(\frac{c_1\varepsilon}{3RJ^3}\right)^2}{3+4\left(\frac{c_1\varepsilon}{3RJ^3}\right)}\right\} \end{aligned}$$

For I_2, we have

$$P\left(\left|I_2\right| > \frac{\varepsilon}{3}\right) \le \sum_{j}^{J_k} P\left(\left|\hat{w}_{k,j} - w_{k,j}\right| > \frac{c_1\varepsilon}{3J^3}\right) \le J^3 2\exp\left\{-\frac{6n\left(\frac{c_1\varepsilon}{3J^3}\right)^2}{3+4\left(\frac{c_1\varepsilon}{3J^3}\right)}\right\}$$

For I_3, we have

$$\begin{aligned} I_3 &= \sum_{j=1}^{J_k}\sum_{r=1}^{R}\left[\left(\hat{w}_{k,j}\hat{p}_{k,jr} + w_{k,j}p_{k,jr}\right)\left(\hat{p}_{k,jr} - p_{k,jr}\right) + \hat{p}_{k,jr}p_{k,jr}\left(w_{k,j} - \hat{w}_{k,j}\right)\right] \\ &= \sum_{j}^{J_k}\sum_{r=1}^{R}\left[\left(\hat{w}_{k,j}\hat{p}_{k,jr} + w_{k,j}p_{k,jr}\right)\left(\hat{p}_{k,jr} - p_{k,jr}\right)\right] + \sum_{j}^{J_k}\sum_{r=1}^{R}\hat{p}_{k,jr}p_{k,jr}\left(\hat{w}_{k,j} - w_{k,j}\right) \\ &:= I_{31} + I_{32} \end{aligned}$$

For I_{31} and I_{32}, we have

$$P\left(\left|I_3\right| > \frac{\varepsilon}{3}\right) \le P\left(\left|I_{31}\right| > \frac{\varepsilon}{6}\right) + P\left(\left|I_{32}\right| > \frac{\varepsilon}{6}\right)$$

$$\begin{aligned} P\left(\left|I_{31}\right| > \frac{\varepsilon}{6}\right) &\le \sum_{j}^{J_k}\sum_{r=1}^{R} P\left(\left|\left(\hat{w}_{k,j}\hat{p}_{k,jr} + w_{k,j}p_{k,jr}\right)\left(\hat{p}_{k,jr} - p_{k,jr}\right)\right| > \frac{\varepsilon}{6}\right) \\ &\le \sum_{j}^{J_k}\sum_{r=1}^{R} p\left(\left|\hat{p}_{k,jr} - p_{k,jr}\right| > \frac{c_1\varepsilon}{6RJ^3}\right) \\ &\le RJ^3 2\exp\left\{-\frac{6n\left(\frac{c_1\varepsilon}{6RJ^3}\right)^2}{3+4\left(\frac{c_1\varepsilon}{6RJ^3}\right)}\right\} \end{aligned}$$

$$\begin{aligned} P\left(\left|I_{32}\right| > \frac{\varepsilon}{6}\right) &\le \sum_{j}^{J_k}\sum_{r=1}^{R} P\left(\left|\hat{p}_{k,jr}p_{k,jr}\left(\hat{w}_{k,j} - w_{k,j}\right)\right| > \frac{\varepsilon}{6}\right) \\ &\le \sum_{j}^{J_k}\sum_{r=1}^{R} P\left(\left|\hat{w}_{k,j} - w_{k,j}\right| \frac{c_1\varepsilon}{6RJ^3}\right) \\ &\le RJ^3 2\exp\left\{-\frac{6n\left(\frac{c_1\varepsilon}{6RJ^3}\right)^2}{3+4\left(\frac{c_1\varepsilon}{6RJ^3}\right)}\right\} \end{aligned}$$

In a word, we have the inequality,

$$P\left(\left|\hat{e}_k - e_k\right| > 2\varepsilon\right) \le O\left(RJ^3\right)\exp\left\{-c_5\frac{n\varepsilon^2}{R^2J^6}\right\}$$

where c_5 represents a positive constant.

Proof of Theorem 3.1. By Conditions (C1) to (C3) and Lemma 3, we can get

$$\begin{aligned}
P\left(D \subseteq \hat{D}\right) &\ge P\left(\left|\hat{e}_k - e_k\right| \le cn^{-\tau}, \forall k \in D\right) \\
&\ge P\left(\max_{1\le k\le P}\left|\hat{e}_k - e_k\right| \le cn^{-\tau}\right) \\
&\ge 1 - \sum_{k=1}^{P} P\left(\max_{1\le k\le P}\left|\hat{e}_k - e_k\right| > cn^{-\tau}\right) \\
&\ge 1 - O\left(RJ^3\right)p\exp\left\{-c_5\frac{c^2n^{1-2\tau}}{R^2J^6}\right\} \\
&\ge 1 - O\left(p\exp\left\{-bn^{1-2\tau-2\varepsilon-2\kappa} + (\varepsilon+\kappa)\log n\right\}\right)
\end{aligned}$$

Where b is a positive constant.

Lemma 4 (Lemma A.2 [19]). For any continuous covariate X_k satisfying conditions (C4) and (C5), let $F_k(y,x)$ be the cumulative distribution function of (Y, X_k) and $\hat{F}_k(y,x)$ be the empirical cumulative distribution function, we have $P\left(\left|\hat{F}_k\left(r,\hat{q}_{k,(j)}\right) - F_k\left(r,q_{k,(j)}\right)\right| > \varepsilon\right) \le c_6\exp\left\{-c_7n^{1-2\rho}\varepsilon^2\right\}$ for any $\varepsilon > 0$, $1 \le r \le R$ and $1 \le j \le J_k$, where c_6 and c_7 are two positive constants.

Lemma 5 (Lemma A.3 [19]). Under (C1), (C4) and (C5), for any $0 < \varepsilon < 1$, so for continuous X_k, we have $P\left(\left|\hat{e}_k - e_k\right| > 2\varepsilon\right) \le O(RJ)\exp\left\{-c_9\frac{n^{1-2\rho}\varepsilon^2}{R^4J^4}\right\}$, there exists a positive constant c_9.

Proof of Theorem 3.2. According to Lemma 4 and Lemma 5, the proof of Theorem 3.2 is the same as Theorem 3.1 and hence is omitted.

Proof of Theorem 3.3. According to Lemma 3 and 5 and under Conditions (C1), (C4), (C5) and (C7). The proof of Theorem 3.2 is proved by [19] and hence is omitted.