Advances in Applied Mathematics 应用数学进展, 2024, 13(3), 967-980
Published Online March 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.133091

Hans汉斯

基于GINI相关系数的超高维分类特征变量筛选

司　萌，张俊英*，张　妍

太原理工大学数学学院，山西　太原

收稿日期：2024年2月25日；录用日期：2024年3月19日；发布日期：2024年3月26日

摘　要

本文提出了一种基于GINI相关系数的超高维判别分析的GINI相关特征筛选方法。该方法在简单条件下建立了稳健筛选性能。首先，对重尾分布、潜在异常值情况下的数据具有稳健性。其次，它没有具体的数据模型限制，适用于参数及非参数模型。第三，由于所得统计量的有界性，在理论推导上比较简单。第四，筛选指标结构简单，计算成本低。通过蒙特卡罗模拟和实际数据实例验证了该方法的有效性。

关键词

距离相关，特征筛选，GINI均值差异，GINI相关系数

Selection of Ultra-High Dimensional Classification Feature Variables Based on GINI Correlation Coefficient

Meng Si, Junying Zhang*, Yan Zhang

Department of Mathematics, Taiyuan University of Technology, Taiyuan Shanxi

Received: Feb. 25th, 2024; accepted: Mar. 19th, 2024; published: Mar. 26th, 2024

Abstract

We proposed a new method named GINI correlation feature screening for ultrahigh dimensional discriminant analysis based GINI correlation coefficients. We also establish the sure screening property for the proposed procedure under simple assumptions. The new procedure has some

*通讯作者。

文章引用: 司萌, 张俊英, 张妍. 基于 GINI 相关系数的超高维分类特征变量筛选[J]. 应用数学进展, 2024, 13(3): 967-980. DOI: 10.12677/aam.2024.133091

additional desirable characters. First, it is robust against heavy-tailed distributions, potential outliers and the sample shortage for some categories. Second, it is model-free without any specification of a regression model and directly applicable to the situation with many categories. Third, it is simple in theoretical derivation due to the boundedness of the resulting statistics. Forth, it is relatively inexpensive in computational cost because of the simple structure of the screening index. Monte Carlo simulations and real data examples are used to demonstrate the finite sample performance.

Keywords

Distance Correlation, Feature Screening, GINI Mean Difference, GINI Correlation Coefficients

1. 引言

变量选择或变量筛选在高维数据分析中起着重要的作用。通过考虑协变量与响应变量之间的关联强度来选择重要变量对于超高维数据是必不可少的，并且在最近的文献中受到了广泛的关注。Fan 和 Lv (2008) [1]提出了线性模型的确定独立筛选(SIS)，它基于对协变量与响应的边际 Pearson 相关性的大小进行排序。此后其他研究者做了大量的工作来将这一过程推广到各种其他类型的模型，包括：广义线性模型[2]、非参数加性模型[3]、Cox 比例风险模型[4]、线性分位数模型[5]和变系数模型(Fan 等 2014 [6]，Zhang 等[7])。不假设任何特定的先验模型的无模型筛选方法也已发展。包括：Li 等人(2012)的基于距离相关的方法[8]，Mai 等人(2015)的融合 Kolmogorov 滤波器[9]，Liu 和 Wang (2017)的条件距离相关方法[10]，Huang 和 Zhu (2016)的基于最大相关的方法[11]，Shao 和 Zhang (2014)的基于鞅差分的方法[12]，Feng 等人(2017)的基于平滑带宽的方法[13]，张等(2018)的边际条件期望变量筛选方法[14]。这些方法的主要理论结果是所谓的“确定筛选属性”，即在适当的条件下，可以将特征空间的维度从 $p_n = o\left(\exp\left(n^{\alpha}\right)\right)$ 降至更小的 $d_n\left(d_n \le n\right)$，同时保留所有相关预测因子的概率接近 1。然而，它是基于连续响应变量的，不能很好地用于分类响应变量。本文旨在研究一种有效的不限定模型的特征筛选方法，用于分类响应变量超高维数据。

本文提出了一种有效的判别分析的确定筛选方法：GINI 超高维特征筛选方法(GCSIS)。该方法在超高维判别分析的背景下，对具有分类响应的数据进行筛选。在不假设预测因子矩条件的情况下，建立了确定的筛选和排序一致性性质。数值研究表明，该方法具有良好的性能。它有以下优点：它是无模型的，因为它的实现不需要指定回归模型；其相应的边际效用可以很容易地评估，而不涉及数值优化；可直接应用于具有分类预测因子的连续响应数据。在全基因组关联研究(GWAS)等实际应用中，该方法尤为有用，其中表型(即响应)是连续的，而单核苷酸多态性(SNPs)等预测因子则是分类的。因此，通过这种方法，可以更有效地筛选出与连续响应相关的重要特征，为科学研究提供有力支持。

本文的其余部分安排如下：第 2 章通过考虑一维数值变量和分类变量之间相关性出发建立与 GINI 均值差异的联系。广义 GINI 相关的性质在 2.1 节中进行了研究。变量筛选方法(GCSIS)在 2.2 节中提出。第 2.3 节是 GCSIS 的理论性质。在第 3 章中，通过仿真和真实数据应用进行了实验研究，以展示 GCSIS 的优势。

2. 筛选方法

2.1. GINI 距离

设$Y\in\{y_1,y_2,\cdots,y_K\}$是分类(K 类)响应变量，$X\in K_X$，其中$K_X$为协变量$X$的支持集合。为了研究$X$和$Y$之间的相关性，本文很自然地考虑给定$Y$和$X$的的条件分布函数，表示为$F(x|Y)=P(X\le x|Y)$。用$F(x)=P(X\le x)$表示$X$的无条件分布函数，用$F_k(x)=P(X\le x|Y=y_k)$表示给定$Y=y_k$，$p_k=P(Y=y_k)$的$X$的条件分布函数，$(X,Y=y_k)$的联合分布是$p_kF_k$，$X$的边际分布是$F(x)=\sum_{k=1}^{K}p_kF_k(x)$。如果对于任何$x\in K_X$且$k=1,\cdots,K$，有$F_k=F(x)$，则$X$和$Y$是独立的。为了衡量$X$和$Y$之间的相关性，Dang 等(2019) [15]考虑了以下条件分布函数和边际分布函数之间的距离相关度量。

$$D:=\mathbb{E}\int_{\mathbb{R}}\left(F(x|Y)-F(x)\right)^2\mathrm{d}x=\sum_{k=1}^{K}p_k\int_{\mathbb{R}}\left(F_k(x)-F(x)\right)^2\mathrm{d}x. \tag{1}$$

显然，当且仅当X和Y独立时，相关性为零。F的 GINI 平均差(GMD) ([16] [17] [18])是

$$\Delta=\Delta(X)=\Delta(F)=\mathbb{E}|X-X'| \tag{2}$$

它表示两个独立随机变量之间的期望距离，其中X和X'是$\mathbb{R}$中具有有限一阶矩的分布F中的独立随机变量。Dorfman (1979) [19]证明了对于非负随机变量，

$$\Delta=2\int F(x)\left(1-F(x)\right)\mathrm{d}x. \tag{3}$$

注意(3)也适用于离散随机变量。因此，相关性可以写成

$$\rho(X,Y)=1-\frac{\sum_{k=1}^{K}p_k\Delta_k}{\Delta}=\frac{\Delta-\sum_{k=1}^{K}p_k\Delta_k}{\Delta}, \tag{4}$$

其中Δ_k是F_k的 GINI 系数(GMD)。$\rho(X,Y)=0$当且仅当X和Y是独立的。

2.2. GINI 相关系数变量筛选方法

本文提出了一种新的无模型确定独立筛选方法，使用$\rho(X,Y)$对超高维定量和定性协变量进行筛选。设Y为具有离散支持$y_1,y_2,\cdots,y_K$ ($K\ge2$)的响应变量，$x=(X_1,\cdots,X_p)^{\mathrm{T}}$为预测向量，其中$p\ge n$且$n$是样本量。在不指定回归模型的情况下，通过

$$M_*=\{j:F(y|X)\text{函数依赖于 }X_j\},$$

定义重要预测变量子集，用$\mathcal{I}=\{1,2,\cdots,p\}\setminus M_*$表示非重要预测变量子集。由第 2.1 节的(4)式知$\rho(X_j,Y)$可以度量X_j与Y之间的相关性。令

$$w_j=\rho(X_j,Y)=\frac{\Delta^j-\sum_{k=1}^{K}p_k\Delta_k^j}{\Delta^j},$$

其中$\Delta^j=E|X_j-X_j'|$，$\Delta_k^j=E\left[|X_j-X_j'||Y=y_k\right]$。

注意到，当部分正交条件(Huang, Horowitz 和 Ma, 2008 [20]；Fan 和 Song, 2010 [2])成立时，即$\{X_j:j\in M_*\}$与$\{X_j:j\in M_*^c\}$独立，当$j\in M_*$时$w_j>0$，当$j\in M_*^c$时$w_j=0$，所以w_j可以作为选择重要变量的一个标准。因此选择重要变量$\{X_j:j\in M_*\}$只和Y的相关性有关，与模型的选择无关。

假设样本数据 $D=\{(\mathbf{x}_i,Y_i)\}$， $i=1,\cdots,n$。设 I_k 为 $Y_i=y_k$ 样本点的索引集，则 p_k 由该类别的样本比例估计，即 $\hat{p}_k=\frac{n_k}{n}$，其中 n_k 为 I_k 中的元素数。下面给出 ρ_g 的一个点估计量。

$$\hat{\Delta}_k^j=\binom{n_k}{2}^{-1}\sum_{i<l\in I_k}\left|X_{i,j}-X_{l,j}\right|,$$

$$\hat{\Delta}^j=\binom{n}{2}^{-1}\sum_{1=i<l=n}\left|X_{i,j}-X_{l,j}\right|,$$

$$\hat{w}_j=\hat{\rho}_j\left(X_j,Y\right)=1-\frac{\sum_{k=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}. \tag{5}$$

基于 $\hat{w}_j$ 本文选择了一组具有较大 $\hat{w}_j$ 的重要预测因子，

$$\hat{M}=\left\{j:1\le j\le p,\hat{w}_j>c_3n^{-r}\right\},$$

其中 $c_3>0$ 且 $0<r<1$。

2.3. 理论性质

显然，$\hat{\Delta}_k^j$ 和 $\hat{\Delta}^j$ 为 U 统计量。应用 U 统计量理论[21] [22]，在以下的假设条件下，可以建立 $\hat{\Delta}_k^j$ 和 $\hat{\Delta}^j$ 的渐近性质。

- **C1.** 矩条件：存在正常数 $M_1,M_2\left(M_1\le M_2\right)$ 和 t_0 使得

$$\max_{1\le j\le p}E\left(\exp\left(tX_j\right)\right)<\infty,0<t<t_0,$$

且

$$M_1\le\left|E\left(X_j\right)\right|\le M_2.$$

- **C2.** 存在 $r\in(0,1)$ 和 $c_3>0$，有 $\min_{j\in M_*}w_j\ge 2c_3n^{-r}$。

定理 2.1 在条件 C1 下，对于任意 $\varepsilon>0$，存在正常数 M_6 和 $s_\varepsilon\in\left(0,2/\left(M_6\varepsilon\right)\right)$，使得

$$P\left\{\max_{1\le j\le p}\left|\hat{w}_j-w_j\right|>\varepsilon\right\}\le 3p\left(1-M_6s_\varepsilon\varepsilon/2\right)^{n/2}.$$

此外，令 $\varepsilon=2/\left(M_6s_\varepsilon\right)-c_1n^{-r}$，$2/\left(M_6s_\varepsilon\right)=2c_1n^{-r}$，假设条件(C2)成立，则

$$P\left(M_*\subset\hat{M}\right)\ge 1-3s_nn\left(1-M_6s_\varepsilon\varepsilon/2\right)^{n/2}.$$

证明：注意到 $\hat{\Delta}_k^j$ 可以表示为如下形式：

$$\begin{aligned}\hat{\Delta}_k^j&=\binom{n_k}{2}^{-1}\sum_{i<l<\mathcal{I}_k}\left[\left(X_{i,j}-X_{l,j}\right)I\left(X_{i,j}>X_{l,j}\right)+\left(X_{l,j}-X_{i,j}\right)I\left(X_{l,j}>X_{i,j}\right)\right]\\&\overset{def}{=}\frac{2}{n_k\left(n_k-1\right)}\sum_{i<l\in\mathcal{I}_k}h_1\left(X_{i,j},X_{l,j}\right).\end{aligned}$$

因此，$\hat{\Delta}_k^j$ 是标准的 U-统计量。利用马尔可夫不等式，可以得到，对任意的 $0<t<s_0j^*$，其中 $j^*=\left[n_k/2\right]$，有

$$P\left(\left|\hat{\Delta}_k^j-\Delta_k^j\right|>\varepsilon\right)\le\exp\{t\varepsilon\}\exp\left\{-t\Delta_k^j\right\}E\left[\exp t\hat{\Delta}_k^j\right].$$

通过 Serfling (1980) [23]的 5.1.6，U-统计 $\hat{\Delta}_k^j$ 可以表示为独立且同分布的随机变量的平均值；即 $\hat{\Delta}_k^j=(n_k!)^{-1}\sum_{n_k!}w\left(X_{1j},\cdots,X_{n_kj}\right)$，其中每个 $w\left(X_{1j},\cdots,X_{n_kj}\right)$ 是 $j^*=[n_k/2]$ 个独立的同分布随机变量的平均值，$n_k!$ 表示 $n_k!$ 种排列 $i_1,\cdots,i_{n_k}$ $(1,\cdots,n_k)$. 记 $\varphi_h(s)=E\left[\exp\left\{sh\left(X_{i,j},X_{l,j}\right)\right\}\right]$，$0<s<s_0$。由于指数函数是凸函数，它遵循 Jensen 不等式

$$\begin{aligned}E\left[\exp\left\{t\Delta_k^j\right\}\right]&=E\left[\exp\left\{t(n_k!)^{-1}\sum_{n_k!}w\left(X_{1j},\cdots,X_{n_kj}\right)\right\}\right]\\&\le(n_k!)^{-1}E\left[\exp\left\{\sum_{n_k!}w\left(X_{1j},\cdots,X_{n_kj}\right)\right\}\right]\\&=\varphi_h^{j^*}\left(t/j^*\right).\end{aligned}$$

结合以上两个结果，有

$$\begin{aligned}P\left(\left|\hat{\Delta}_k^j-\Delta_k^j\right|>\varepsilon\right)&\le\exp\{t\varepsilon\}\left[\exp\left\{-t\Delta_k^j/j^*\right\}\varphi_h\left(t/j^*\right)\right]^{j^*}\\&=\left[\exp\{s\varepsilon\}\exp\left\{s\Delta_k^j\right\}\varphi_h(s)\right]^{j^*},\end{aligned}$$

其中 $s=t/j^*$。注意到 $E\left(h_1\left(X_{i,j},X_{l,j}\right)\right)=\Delta_k^j$，且对于任意随机变量 Y，泰勒展开式 $\exp\{sY\}=1+sY+s^2Z/2$，其中 $0<Z<Y^2\exp\{s_1Y\}$，且 s_1 是介于 0 和 s 之间的常数。因此

$$\exp\left\{s\Delta_k^j\right\}\varphi_h(s)\le1+s^2\left[Eh_1^4\left(X_{i,j},X_{l,j}\right)\times E\exp\left\{2s_1\left(h_1-\Delta_k^j\right)\right\}\right]^{1/2}\Big/2. \tag{6}$$

由条件 C1 可得存在常数 C (n 和 p 的独立性)使得 $\exp\left\{-s\Delta_k^j\right\}\varphi_h(s)\le Cs^2$，即

$$\exp\left\{-s\Delta_k^j\right\}\varphi_h(s)=1+O\left(s^2\right).$$

对于充分小的 s，可以通过选择足够小的 t 来实现，有 $\exp(-s\varepsilon)=1-\varepsilon s+O\left(s^2\right)$，因此

$$\exp(-s\varepsilon)\exp\left(-s\Delta_k^j\right)\varphi_h(s)\le1-s\varepsilon/2. \tag{7}$$

结合结果(6)，(7)和 $p_k=\hat{p}_k=\frac{n_k}{n}$，可以得到对任意 $\varepsilon>0$，存在一个充分小的 s_ε，使得 $P\left(\left|\Sigma_{i=1}^K\hat{p}_k\hat{\Delta}_k^j-\Sigma_{i=1}^Kp_k\Delta_k^j\right|>\varepsilon\right)\le\left(1-s_\varepsilon\varepsilon/2\right)^{n/2}$。这里用符号 s_ε 来强调 s 依赖于 ε。同理，可以证明对于任何 $\varepsilon_1>0$，$P\left(\left|\hat{\Delta}^j-\Delta^j\right|>\varepsilon_1\right)\le\left(1-s_{\varepsilon_1}\varepsilon_1/2\right)^{n/2}$。

下面考虑 $P\left(\left|\frac{\Sigma_{i=1}^K\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^Kp_k\Delta_k^j}{\Delta^j}\right|\ge\varepsilon\right)$。首先证明 $\hat{\Delta}^j$ 离 0 有统一的边界，其概率趋于 1。由条件 C1，存在 $M_3>0$ 使得 $\left|\Delta^j\right|\ge M_4>0$ 且对于 $\delta_0\in(0,1)$ 使得 $M_4\equiv M_3-\delta_0$。则由(7)可知对某个正常数 s_{δ_0}，有

$$\begin{aligned}P\left(\left|\hat{\Delta}^j\right|\le M_4\right)&\le P\left(\left|\Delta^j\right|-\left|\hat{\Delta}^j-\Delta^j\right|\le M_3-\delta_0\right)\\&\le P\left(\left|\hat{\Delta}^j-\Delta^j\right|\ge\delta_0\right)\\&\le\left(1-s_{\delta_0}\delta_0/2\right)^{n/2}.\end{aligned}$$

因此

$$
\begin{aligned}
&P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon\right)\\
&=P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon,\left|\hat{\Delta}^j\right|\leq M_4\right)+P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon,\left|\hat{\Delta}^j\right|>M_4\right)\\
&\leq P\left(\left|\hat{\Delta}^j\right|\leq M_3\right)+P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon,\left|\hat{\Delta}^j\right|>M_4\right)\\
&\leq\left(1-s_{\delta_0}\delta_0/2\right)^{n/2}+P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon,\left|\hat{\Delta}^j\right|>M_4\right).
\end{aligned}
\tag{8}
$$

(8)的第二项是

$$
\begin{aligned}
&P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon,\left|\hat{\Delta}^j\right|>M_4\right)\\
&\leq P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\hat{\Delta}^j}\right|+\left|\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon,\left|\hat{\Delta}^j\right|>M_4\right)\\
&\leq P\left(\frac{1}{M_3}\left|\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j-\Sigma_{i=1}^{K}p_k\Delta_k^j\right|\geq\varepsilon/2\right)+P\left(\frac{\left|\Sigma_{i=1}^{K}p_k\Delta_k^j\right|}{\left|\hat{\Delta}^j\right|\left|\Delta_k^j\right|}\left|\hat{\Delta}^j-\Delta^j\right|\geq\varepsilon/2\right)\\
&\leq\left(1-s_\varepsilon M_4\varepsilon/2\right)^{n/2}+\left(1-s_{\varepsilon_1}M_3M_4M_5\varepsilon_1/2\right)^{n/2},
\end{aligned}
\tag{9}
$$

其中 $M_5>0$。从条件 C1 可以知道 $\left|\Sigma_{i=1}^{K}p_k\Delta_k^j\right|$ 有界，并假设 $\left|\Sigma_{i=1}^{K}p_k\Delta_k^j\right|\leq M_5$。

因此(7)简化为

$$
\begin{aligned}
&P\left(\left|\frac{\Sigma_{i=1}^{K}\hat{p}_k\hat{\Delta}_k^j}{\hat{\Delta}^j}-\frac{\Sigma_{i=1}^{K}p_k\Delta_k^j}{\Delta^j}\right|\geq\varepsilon\right)\\
&\leq\left(1-s_{\delta_0}\delta_0/2\right)^{n/2}+\left(1-s_\varepsilon M_4\varepsilon/2\right)^{n/2}+\left(1-s_{\varepsilon_1}M_3M_4M_5\varepsilon_1/2\right)^{n/2}\\
&\leq 3\left(1-M_6s_\varepsilon\varepsilon/2\right)^{n/2},
\end{aligned}
\tag{10}
$$

其中 $M_6=\min\left\{\frac{s_{\delta_0}\delta_0}{s_\varepsilon\varepsilon},M_4,\frac{s_{\varepsilon_1}M_3M_4M_5}{s_\varepsilon\varepsilon}\right\}$。这就完成了定理 2.1 第一部分的证明。

进一步在条件(C2)下，证明

$$P\left(M_*\subset\hat{M}\right)\geq 1-3s_n n\left(1-M_6s_\varepsilon\varepsilon/2\right)^{n/2}.$$

根据 $\hat{M}$ 的定义和条件 C2，

$$
\begin{aligned}
P\left(M_*\subset\hat{M}\right)&=P\left(\min\nolimits_{j\in M_*}\hat{w}_j>c_3n^{-r}\right)=P\left(\min\nolimits_{j\in M_*}w_j-\min\nolimits_{j\in M_*}\hat{w}_j\leq\min\nolimits_{j\in M_*}w_j-c_3n^{-r}\right)\\
&\geq P\left(\min\nolimits_{j\in M_*}w_j-\min\nolimits_{j\in M_*}\hat{w}_j\leq 2c_3n^{-r}-c_3n^{-r}\right)\geq P\left(\max\nolimits_{j\in M_*}\left|w_j-\hat{w}_j\right|\leq c_3n^{-r}\right)\\
&=1-P\left(\max\nolimits_{j\in M_*}\left|w_j-\hat{w}_j\right|\geq c_3n^{-r}\right)\geq 1-s_n\max\nolimits_{j\in M_*}P\left(\left|w_j-\hat{w}_j\right|\geq c_3n^{-r}\right)\\
&\geq 1-3s_nn\left(1-M_6s_\varepsilon\varepsilon/2\right)^{n/2}.
\end{aligned}
$$

这就完成了定理 2.1 第二部分的证明。

3. 数值研究

在本节中，首先通过蒙特卡罗模拟研究评估提出的 GINI 相关未来筛选(GCSIS)的有限样本性能。然后，通过两个真实数据实例进行实证分析，以说明所提出的 GCSIS 方法的有效性。

3.1. 模拟研究

本文使用包括所有重要变量的最小模型大小(MMS)来衡量每种筛选方法的效果。另外，对于给定的模型大小 $d=\lceil n/\log n\rceil$，其中 n 为样本量，$\lceil x\rceil$ 为 x 的整数部分。用 T_j 表示包含单个重要变量 X_j 的比例，用 M_a 表示包含所有重要变量的比例。所有数值研究都是使用 R 代码进行的。

例 3.1 (超高维线性判别分析(Cui, *et al.*) [24])在这个例子中，本文考虑一个具有超高维预测变量的线性判别分析问题，通过遵循 Pan，Wang 和 Li (2013) [25]中的类似设置。对于每个第 i 次的观测样本，类别响应 Y_i 由两个不同的分布生成：1) 均衡的，有 K 个类别的离散均匀分布，其中 $P(Y_i=k)=1/K$，$k=1,\cdots,K$；2) 非均衡的，概率序列 $p_k=P(Y_i=k)=2\left[1+(k-1)/(K-1)\right]/3K$ 是一个等差数列 $\max_{1\le j\le p}p_k=2\min_{1\le j\le p}p_k$。例如，当 Y 是二进制时，$p_1=1/3$，$p_2=2/3$。给定 $Y_i=k$，则通过令 $X_i=u_k+\varepsilon_i$ 生成第 i 个预测变量 X_i，其中平均项 $u_k=(u_{k1},\cdots,u_{kp})\in K^p$ 是 p 维向量，第 k 个分量 $u_{kk}=3$，而其他分量均为零，$\varepsilon_i=(\varepsilon_{i1},\cdots,\varepsilon_{ip})$ 是一个 p 维误差项。这里，考虑两种情况下的误差项：1) $\varepsilon_{ij}\sim N(0,1)$；2) $\varepsilon_{ij}\sim t(2)$ 分别对每个 $j=1,\cdots,p$ 成立。请注意，情况 2)使每个预测变量都具有重尾，其目的是检查独立筛选方法的稳健性。为了系统地检查 GSIS 和其他方法，考虑 2000 个预测因子和 $n=40$ 的二元响应变量，以及每种情况下 $n=200$ 的 10 分类响应。即分别取 $(R,n,p)=(2,40,2000)$ 和 $(10,200,2000)$。

Table 1. Simulation results of linear discriminant analysis, $R=2$ (example 3.1)
表 1. 线性判别分析的仿真结果，$R=2$ (例 3.1)

Pr	方法	情况(1)：$\varepsilon_{ij}\sim N(0,1)$				情况(2)：$\varepsilon_{ij}\sim t(2)$			
		MMS	T_1	T_2	M_a	MMS	T_1	T_2	M_a
均衡	SIS	2.0 (0.0)	1.00	1.00	1.00	2.4 (9.8)	0.80	0.90	0.78
	SIRS	2.0 (0.0)	1.00	1.00	1.00	8.5 (23.2)	0.69	0.72	0.50
	DC-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	0.99	0.98	0.97
	KF	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	0.99	0.99	0.98
	PSIS	2.0 (0.0)	1.00	1.00	1.00	2.5 (9.3)	0.80	0.85	0.75
	MV-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	1.00	0.99	0.99
	GCSIS	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	1.00	0.99	0.99
非均衡	SIS	2.0 (0.0)	1.00	1.00	1.00	5.5 (49.0)	0.75	0.75	0.60
	SIRS	2.0 (0.0)	1.00	0.99	0.99	17.5 (125.3)	0.70	0.64	0.45
	DC-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	0.95	0.096	0.90
	KF	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	0.96	0.98	0.95
	PSIS	2.0 (0.0)	1.00	1.00	1.00	6.0 (49.0)	0.75	0.75	0.56
	MV-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	0.96	0.99	0.95
	GCSIS	2.0 (0.0)	1.00	1.00	1.00	2.0 (0.0)	0.98	0.99	0.97

Table 2. Simulation results of linear discriminant analysis, R = 10 (example 3.1)
表 2. 线性判别分析的仿真结果，R = 10 (例 3.1)

方法	MMS	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	M_a
均衡概率和情况(1)：$\varepsilon_{ij} \sim N(0,1)$												
DC-SIS	10.0 (0.0)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.99
PSIS	10.0 (0.0)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MV-SIS	10.0 (0.0)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
GCSIS	10.0 (0.0)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
均衡概率和情况(2)：$\varepsilon_{ij} \sim t(2)$												
DC-SIS	16.0 (22.5)	0.85	0.96	0.98	0.98	0.97	0.98	0.99	0.99	0.99	0.98	0.75
PSIS	371.2 (570.3)	0.74	0.75	0.76	0.74	0.75	0.75	0.73	0.76	0.76	0.80	0.06
MV-SIS	11.3 (4.0)	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.95
GCSIS	10.5 (3.5)	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.97
非均衡概率和情况(1)：$\varepsilon_{ij} \sim N(0,1)$												
DC-SIS	13.2 (15.0)	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85
PSIS	10.0 (0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MV-SIS	10.0 (0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GCSIS	10.0 (0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
非均衡概率和情况(2)：$\varepsilon_{ij} \sim t(2)$												
DC-SIS	130.1 (248.8)	0.35	0.91	0.93	0.93	0.97	1.00	0.99	1.00	1.00	1.00	0.23
PSIS	350.5 (446.7)	0.68	0.66	0.56	0.58	0.64	0.63	0.62	0.60	0.73	0.62	0.07
MV-SIS	13.0 (10.0)	0.94	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.81
GCSIS	12.3 (10.2)	0.95	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.85

首先，比较了 GSIS 与 MV-SIS (Cui, *et al*., 2015) [24]、SIS (Fan 和 Lv, 2008) [1]、SIRS (Zhu, *et al*., 2011) [26]、DC-SIS (Li, Zhong 和 Zhu, 2012) [8]、Kolmogorov Filter (Mai 和 Zou 等, 2015) [9]和 PSIS (Pan, Wang 和 Li, 2013) [25]在二元响应变量的性能，其中X_1和X_2是重要变量。表 1 总结了在给定模型大小$d = \lfloor n/\log n \rfloor$下，每种方法基于 500 次模拟的 MMS 的中位数及其相关的标准偏差的稳健估计(RSD = IQR/1.34)，括号中的为T_j，$j = 1, 2, T$。

接下来，考虑 10 个类别的响应，其中$X_1, X_2, \cdots, X_{10}$是重要变量。注意到，响应变量 Y 是一个名义数字，这使得 SIS、SIRS 和 Kolmogorov Filter 不适用。为了使 DC-SIS 适用于该问题，将 10 个分类响应转换为 9 个虚拟二元变量，定义为一个新的多重响应变量。注意到，Li，Zhong 和 Zhu (2012) [8]认为 DC-SIS 可以用于多重响应。Pan, Wang 和 Li (2013) [25]提出了一种成对确定独立筛选(PSIS)来处理分类反应。PSIS 每次对每一对类(r_j, r_2)使用$|u_{r1j} - u_{r2j}|$作为预测变量X_j的边际信号，其中u_{rj}表示X_{ij}对于$i \in \{i : Y_r = r\}$的样本平均值。表 2 总结了在给定模型大小$d = \lfloor n/\log n \rfloor$下，基于 500 次模拟的 MMS 的中位数及其相关的

稳健标准偏差(括号中为 $j=1,2,\cdots,10$)。

表 1 和表 2 都表明，在线性判别分析中，本文提出的 GCSIS 在变量筛选方面优于其他比较方法。当误差项是重尾的并且响应类别的数量增加时，GCSIS 的最小模型大小(MMS)要小得多，并且与其他独立筛选相比，GCSIS 在所选模型中包含所有重要变量的概率要高得多。因此，GCSIS 的鲁棒性是一个重要的特征，使得它在实践中更有用。当误差项为正态时，GCSIS 的估计和预测性能与 PSIS 非常接近。然而，当误差偏离正态分布时，PSIS 会恶化，而 GCSIS 仍然表现良好。

例 3.2 为了模拟等位基因频率相等的 SNP，本文用 Z_{ij} 表示第 j 个 SNP 对第 i 个受试者的优势效应，并以如下方式生成：如果 $X_{ij}<q_1$， $Z_{ij}=1$；如果 $q_1<X_{ij}<q_3$， $Z_{ij}=0$；如果 $X_{ij}>q_3$， $Z_{ij}=-1$，其中 $X_i=\left(X_{i1},X_{i1},\cdots,X_{ip}\right)\sim N\left(0,\Sigma\right)$， $\Sigma=\rho_{ij}^{|i-j|}$ 其中 $\left(\rho_{ij}\right)_{p\times p},i=1,\cdots,n;j=1,\cdots,p$。$q_1$ 和 q_3 分别是标准正态分布的第一个和第三个四分位数。然后，通过以下方式生成响应变量(某些特征或疾病)：

$$Y=\beta_1Z_1+\beta_2Z_2+2\beta_3Z_{10}+2\beta_4Z_{20}-2\beta_5\left|Z_{100}\right|+\varepsilon,$$

其中 $\beta_j=(-1)^U\left(a+|Z|\right)$， $j=1,\cdots,5$，其中 $a=2\log\left(\frac{n}{\sqrt{n}}\right)$， $U\sim Bernoulli(0.4)$ 且 $Z\sim N(0,1)$，误差项 ε 在 $N(0,1)$ 或 $t(1)$ 之后。

Table 3. Simulation results for example 3.2
表 3. 例 3.2 的模拟结果

ε	方法	MMS	T_1	T_2	T_{10}	T_{20}	T_{100}	M_a
$N(0,1)$	SIS	1120.0 (834.1)	0.96	0.97	1.00	0.99	0.02	0.02
	DCSIS	10.0 (40.3)	0.94	0.95	1.00	0.99	0.80	0.73
	SIRS	1317 (907.3)	0.96	0.95	1.0	0.98	0.03	0.02
	RRCS	1102.0 (810.3)	0.96	0.94	0.99	0.98	0.03	0.03
	MV-SIS	8.0 (35.2)	0.96	0.94	0.99	0.98	0.89	0.78
	GCSIS	8.6 (31.5)	0.96	0.95	0.99	0.98	0.90	0.82
$t(1)$	SIS	1430.0 (550.1)	0.30	0.35	0.42	0.42	0.02	0.00
	DCSIS	125.0 (290.3)	0.78	0.76	0.94	0.91	0.53	0.33
	SIRS	1126.0 (625.1)	0.86	0.84	0.96	0.96	0.02	0.01
	RRCS	1024.0 (730.1)	0.87	0.84	0.98	0.96	0.02	0.01
	MV-SIS	46.2 (141.3)	0.79	0.79	0.94	0.94	0.79	0.47
	GCSIS	38.6 (180.3)	0.80	0.80	0.92	0.92	0.80	0.50

有 5 个有效的 SNP，分别是 Z_1,Z_2,Z_{10},Z_{20} 和 Z_{100}。前 4 个活性 SNP 与响应 Y 呈线性相关，而 SNP Z_{100} 与 Y 呈非线性相关。有趣的是，显性效应的绝对值 $\left|Z_{100}\right|$ 是遗传学中对应的加成效应。在这里，考虑 5 种不同的独立筛选方法：MV-SIS、SIS、DC-SIS、SIRS、RRCS (Li *et al*., 2012 [8])和 GCSIS，并设置 $n=200$ 和 $p=2000$，每个实验重复 500 次。在表 3 中总结了 $d=\left\lceil n/\log(n)\right\rceil$ 的模拟结果。

由表 3 可知，当误差服从正态分布时，由于与响应呈线性相关关系，5 种独立筛选都能有效地筛选

出前 4 个活性 SNP。然而，只有 DC-SIS、MV-SIS 和 GCSIS 可以选择对 Y 有非线性贡献的 Z_{100}。当误差由很大程度上是重尾的 $t(1)$ 产生时，所有独立筛选方法的表现都不如以前。然而，GCSIS 的性能仍然是最好的。由此，可以得出 GCSIS 可以有效地选择与响应变量线性或非线性相关的活性分类 SNP。

例 3.3 (非参数可加模型)根据 Meier，Geer 和 Buhlmann (2009) [27]，本文定义了以下四个函数

$$f_1(x)=-\sin(2x),\ f_2(x)=x^2-12/25,\ f_3(x)=x,\ f_4(x)=\mathrm{e}^{-x}-2/5\sinh(5/2).$$

然后考虑以下加性模型：

$$Y=3f_1(X_1)+f_2(X_2)+1.5f_3(X_3)+f_4(X_4)+\varepsilon,$$

其中预测变量独立于 $U[-2.5,2.5]$ 生成。为了检验每种独立筛选方法的鲁棒性，本文考虑误差项 $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)$ 的两种情况：1) $\varepsilon_i\sim N(0,1)$；2) $\varepsilon_i\sim t(1)$，$i=1,2,\cdots,n$。设 n = 200，p = 2000，每个错误情况下每个实验重复 500 次。在我们的模拟中，使用第一、第二和第三、四分位数作为 GSIS 的节，将每个预测器离散为一个 4 分类变量。表 4 中报告了给定模型尺寸 $d=\lceil n/\log(n)\rceil$ 的模拟结果。

表 4 表明，GCSIS 在离散化每个预测因子后表现非常好。虽然 DCSIS 可能会检测到非线性，但它偶尔会遗漏 X_1 和 X_2。可能的原因是 Y 与前两个预测因子之间的距离相关性相对较弱。另一方面，GCSIS 仍然可以有效地选择主动预测因子，这再次显示了它的鲁棒性。

Table 4. Simulation results for example 3.3
表 4. 例 3.3 的模拟结果

ε	方法	MMS	T_1	T_2	T_3	T_4	M_a
$N(0,1)$	SIS	1203.5 (721.5)	0.16	0.02	1.00	1.00	0.00
	DCSIS	50.5 (56.2)	0.47	0.79	1.00	1.00	0.35
	SIRS	1221.0 (670.3)	0.16	0.01	1.00	1.00	0.90
	RRCS	1143.0 (682.1)	0.17	0.03	1.00	1.00	0.00
	MV-SIS	4.0 (1.5)	0.99	0.95	1.00	1.00	0.95
	GCSIS	4.0 (1.8)	0.99	0.96	1.00	1.00	0.96
$t(1)$	SIS	1510.3 (540.2)	0.04	0.01	0.44	0.51	0.00
	DCSIS	205.5 (265.3)	0.20	0.33	0.96	0.96	0.13
	SIRS	1223.5 (658.3)	0.12	0.01	1.00	1.00	0.00
	RRCS	1230.5 (690.1)	0.14	0.01	0.99	1.00	0.00
	MV-SIS	11.0 (25.2)	0.93	0.82	0.99	1.00	0.75
	GCSIS	10.5 (23.5)	0.94	0.82	0.99	1.00	0.78

3.2. 真实数据示例

3.2.1. 肺癌数据

Gordon 等人(2002) [28]和 Fan 和 Fan (2008) [29]先前对肺癌数据进行了分析，以区分肺恶性胸膜间皮瘤(MPM)和肺腺癌(ADCA)。共有来自两类的 12。533 个基因和 181 个组织样本：MPM 类 31 个，ADCA 类 150 个。训练数据集包含 32 个样本(16 个 MPM 和 16 个 ADCA)，而剩下的 149 个样本(15 个 MPM 和

134 个 ADCA)用于测试。

首先将数据标准化为零均值和单位方差。Fan 和 Fan (2008) [29]表明，他们的特征退火独立规则(FAIR)选择了 31 个重要基因，没有产生训练误差和 7 个测试误差，而 Tibshiran 等(2002) [30]提出的最近萎缩质心(NSC)方法选择了 26 个基因，没有产生训练误差和 11 个测试误差。然后，考虑 MV-SIS，DC-SIS，PSIS 和我们的 GCSIS 方法(用 GCSIS1 表示)，使用 LDA 来解决这个超高维分类问题。注意到 FAIR 在 t 检验筛选后使用了对角线性判别分析(LDA)。为了进行公平的比较，还增加了一个将 t 检验筛选与 LDA 相结合的方法，用 FAIR*表示，MV-SIS 后接 LDA (即 MV-SIS1)，MV-SIS 后接 SDA (即 MV-SIS2)。本例还采用了 Witten 和 Tibshirani (2011) [31]提出的惩罚 LDA 方法(用 PenLDA 表示)和 Clemmensen 等(2011) [32]提出的稀疏判别分析(用 SDA 表示进行比较)。此外，本文将 GCSIS 与 SDA 结合起来，认为这种两阶段方法是另一种潜在的方法，用 GCSIS2 表示。

为了评估预测性能，将所有 181 个组织样本随机划分为两部分：包括 100 个样本的训练集和其余 81 个样本的测试集。将上述过程应用于训练数据，并通过训练集和测试集的分类误差来评估它们的性能。为了公平的比较，本文使用相同的 BIC 标准为所有方法选择最佳的模型大小。重复实验 100 次，表 5 中总结了训练和测试分类误差与被选基因数的均值及其相关标准差(括号内)。结果表示，GCSIS1 表现相当好，平均使用 12 个左右的基因，训练和测试误差都很小。其中，SDA 方法对训练样本分类效果较好，测试错误率较小。然而，SDA 倾向于选择相当多的基因，因此可能会失去一些模型的可解释性。值得注意的是，GCSIS2 可以用更少的基因数量实现最小的测试错误率。这进一步证明了将 GCSIS 与 SDA 相结合的两阶段方法的优点。

Table 5. Performance evaluation of Lung Cancer Data
表 5. 肺癌数据的性能评估

方法	训练误差(%)	测试误差(%)	被选基因数
NSC	0.87 (0.90)	1.86 (1.93)	17.56 (12.10)
FAIR	3.05 (1.4)	3.52 (2.01)	13.72 (7.40)
PenLDA	0.88 (0.92)	1.95 (1.97)	18.95 (18.14)
SDA	0.00 (0.00)	1.42 (1.21)	39.83 (2.85)
PSIS	0.06 (0.24)	2.14 (1.57)	24.69 (6.85)
DCSIS	0.08 (0.27)	2.63 (2.30)	15.54 (12.53)
MVSIS1	0.15 (0.44)	1.77 (1.91)	11.99 (9.53)
MVSIS2	0.16 (0.43)	2.20 (1.89)	12.42 (9.20)
GCSIS1	0.19 (0.41)	1.61 (1.70)	11.70 (8.35)
GCSIS2	0.18 (0.42)	1.32 (1.50)	11.54 (7.21)

3.2.2. 肺癌数据

该人类肺癌数据是通过 mRNA 表达谱分析的(Bhattacharjee, *et al.*, 2001) [33]。203 例快速冷冻肺肿瘤和正常肺的 mRNA 表达量为 12,600 个。203 个标本被分为 5 个亚类：肺腺癌(ADEN) 139 个，鳞状细胞肺癌(SQUA) 21 个，小细胞肺癌(SCLC) 6 个，肺类癌(COID) 20 个，其余 17 个正常肺样本(normal)。在分

类之前，首先将数据标准化到零均值和单位方差。为了评估所提出方法的预测性能，随机从每个子类中选择大约$100\tau\%$的观测值作为训练样本，其余$100(1-\tau)\%$的观测值作为测试样本，其中$\tau\in(0,1)$。

注意到，前面提到的 NSC 和 FAIR 仅针对二元分类问题提出，因此它们不适用于这种多类判别分析。将带有 LDA 的 psi、DC-SIS、MV-SIS 和 GCSIS 应用于训练集，并通过测试样本对其性能进行评价。对于 DC-SIS、MV-SIS (表示为 MV-SIS1)和 GCSIS (表示为 GSIS1)，采用 LDA 方法，采用留一交叉验证方法为训练数据选择最优模型大小。此外，还考虑了惩罚 LDA (PenLDA)，MV-SIS，然后 SDA (MV-SIS2)和 GCSIS (GCSIS2)进行比较，并使用 10 折叠交叉验证而不是留一交叉验证来选择最佳模型大小，以减少计算时间。虽然 SDA 可以直接应用于给定模型尺寸的多类判别分析，但对于多类超高维数据，为 SDA 寻找最佳模型尺寸的计算成本非常高。因此，使用 GCSIS 降维，然后使用 SDA (即 GSIS2)，而不是在示例中单独使用 SDA。

接下来，本文选择的值为：$\tau=0.9,0.8$，分别重复实验 100 次。根据前面的示例(第 3.2.1 节)，训练和测试分类误差的方法以及所选基因的相应数量及其相关的标准差(在表 6 中报告了括号内的数据)。可以清楚地观察到，虽然所有方法在肿瘤分类中都表现得相当好，但在训练和测试分类误差以及选择基因的数量方面，LDA 或 SDA 的 GCSIS 方法都明显优于其他方法。具体来说，就是 GCSIS+SDA(即 GSIS2)方法利用少量顶级基因实现最佳性能。此外，可以发现 GCSIS 选择的顶级基因不是正态分布的，并且存在潜在的异常值。这一观察结果解释了为什么其他方法的性能相对较差，并证实了所提出的 GCSIS 的鲁棒性特征。该实例进一步证明了将 GCSIS 方法与判别分析相结合的两阶段方法在实际中更有利于超高维数据的处理。

Table 6. Performance evaluation of Lung Cancer Data
表 6. 肺癌数据的性能评价

τ	方法	训练误差(%)	测试误差(%)	被选基因数
0.9	PenLDA	21.88 (2.24)	21.71 (3.87)	25.76 (21.04)
	PSIS	3.54 (0.79)	9.43 (5.65)	107.54 (15.71)
	DC-SIS	6.85 (1.35)	11.81 (6.40)	32.08 (3.85)
	MVSIS1	3.65 (1.15)	7.62 (5.09)	31.76 (10.24)
	MVSIS2	3.60 (1.15)	7.70 (4.91)	31.90 (10.10)
	GCSIS1	3.52 (1.13)	7.64 (5.01)	28.43 (9.86)
	GCSIS2	3.31 (1.10)	7.32 (4.91)	20.71 (80.23)
0.8	PenLDA	22.12 (2.10)	22.40 (4.37)	25.04 (21.81)
	PSIS	3.08 (1.11)	7.90 (3.89)	101.88 (15.72)
	DC-SIS	6.33 (2.16)	13.15 (5.32)	32.18 (5.38)
	MVSIS1	3.74 (2.09)	6.70 (4.24)	27.20 (9.11)
	MVSIS2	3.80 (1.99)	6.65 (4.40)	26.54 (9.04)
	GCSIS1	3.61 (2.12)	6.45 (4.31)	25.94 (8.78)
	GCSIS2	3.23 (2.53)	6.2 (4.10)	25.30 (8.21)

4. 结论与展望

本文提出了一种新的基于 GINI 相关系数的超高维判别分析方法，创新性地提出了 GINI 相关特征筛选策略。此方法不仅在重尾分布和潜在异常值存在的场景下展现出良好的稳健性，更值得一提的是，它没有特定的模型限制，因此可以灵活应用于各种参数和非参数模型。此外，对于非正态分布的数据，该方法同样能够准确识别出关键变量，这在实际应用中具有重要意义。然而，该方法也有其局限性。在处理固定线性模型时，其效果可能不如确定独立筛选(SIS)方法。未来，我们将致力于进一步优化这一方法，以使其在更多场景下都能达到理想的效果。

参考文献

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

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

[3] Fan, J., 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

[4] Fan, J., Feng, Y. and Wu, Y. (2010) High-Dimensional Variable Selection for Cox's Proportional Hazards Model. In: Berger, J.O., Cai, T.T. and Johnstone, I.M., Eds., *IMS Collections, Borrowing Strength*: *Theory Powering Applications, A Festschrift for Lawrence D. Brown*, Vol. 6, IMS, Beachwood, 70-86. https://doi.org/10.1214/10-IMSCOLL606

[5] Ma, S., Li, R. and Tsai, C.-L. (2017) Variable Screening via Quantile Partial Correlation. *Journal of the American Statistical Association*, **112**, 650-663. https://doi.org/10.1080/01621459.2016.1156545

[6] Fan, J., Ma, J. 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

[7] Zhang, J.Y., Zhang, R.Q. and Lu, Z.P. (2016) Quantile-Adaptive Variable Screening in Ultra-High Dimensional Varying Coefficient Models. *Journal of Applied Statistics*, **43**, 643-654. https://doi.org/10.1080/02664763.2015.1072141

[8] Li, R., Zhong, W. and Zhu, L. (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] Mai, Q., Zou, H., *et al*. (2015) The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method. *The Annals of Statistics*, **43**, 1471-1497. https://doi.org/10.1214/14-AOS1303

[10] Liu, Y. and Wang, Q. (2017) Model-Free Feature Screening for Ultrahigh-Dimensional Data Conditional on Some Variables. *Annals of the Institute of Statistical Mathematics*, **23**, 1-19.

[11] Huang, Q. and Zhu, Y. (2016) Model-Free Sure Screening via Maximum Correlation. *Journal of Multivariate Analysis*, **148**, 89-106. https://doi.org/10.1016/j.jmva.2016.02.014

[12] Shao, X. and Zhang, J. (2014) Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening. *Journal of the American Statistical Association*, **109**, 1302-1318. https://doi.org/10.1080/01621459.2014.887012

[13] Feng, Y., Wu, Y. and Stefanski, L.A. (2018) Nonparametric Independence Screening via Favored Smoothing Bandwidth. *Journal of Statistical Planning and Inference*, **197**, 1-14. https://doi.org/10.1016/j.jspi.2017.11.006

[14] Zhang, J.Y., Zhang, R.Q. and Zhang, J.J. (2018) Feature Screening for Nonparametric and Semiparametric Models with Ultrahigh-Dimensional Covariates. *Journal of Systems Science and Complexity*, **31**, 1350-1361. https://doi.org/10.1007/s11424-017-6310-6

[15] Dang, X., Nguyena, D., Chen, Y.X. and Zhang, J.Y. (2019) New Gini Correlation between Quantitative and Qualitative Variables, *Scandinavian Journal of Statistics*, **48**, 1314-1343. https://doi.org/10.1111/sjos.12490

[16] David, H.A. (1968) Gini's Mean Difference Rediscovered. *Biometrika*, **55**, 573-575. https://doi.org/10.2307/2334264

[17] Gini, C. (1914) Sulla misura della concentrazione e della variabiliaà dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Aeti*, **62**, 1203-1248.

[18] Yitzhaki, S. and Schechtman, E. (2013) The Gini Methodology. Springer, New York. https://doi.org/10.1007/978-1-4614-4720-7

[19] Dorfman, R. (1979) A Formula for the Gini Coefficient. *Review of Economics and Statistics*, **61**, 146-149. https://doi.org/10.2307/1924845

[20] Huang, J., Horowitz, J. and Ma, S. (2008) Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models. *The Annals of Statistics*, **36**, 587-613. https://doi.org/10.1214/009053607000000875

[21] Hoeffding, W. (1948) A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, **19**, 293-325. https://doi.org/10.1214/aoms/1177730196

[22] Schechtman, E. (1991) On Estimating the Asymptotic Variance of a Function of U Statistics. *The American Statistician*, **45**, 103-106. https://doi.org/10.2307/2684368

[23] Serfling, R.J. (2009) Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Hoboken.

[24] Cui, H., Li, R. and Zhong, W. (2012) 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

[25] Pan, R., Wang, H. and Li, R. (2013) On the Ultrahigh Dimensional Linear Discriminant Analysis Problem with a Diverging Number of Classes.

[26] Zhu, L.P., Li, L., Li, R. 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

[27] Meier, L., Van De Geer, S. and Bühlmann, P. (2009) High-Dimensional Additive Modeling. *Annals of Statistics*, **37**, 3779-3821. https://doi.org/10.1214/09-AOS692

[28] Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D. and Bueno, R. (2002) Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. *Cancer Research*, **62**, 4963-4967.

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

[30] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *Proceedings of the National Academy of Sciences*, **99**, 6567-6572. https://doi.org/10.1073/pnas.082099299

[31] Witten, D.M. and Tibshirani, R. (2011) Penalized Classification Using Fisher's Linear Discriminant. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **73**, 753-772. https://doi.org/10.1111/j.1467-9868.2011.00783.x

[32] Clemmensen, L., Hastie, T., Witten, D., *et al*. (2011) Sparse Discriminant Analysis. *Technometrics*, **53**, 406-413. https://doi.org/10.1198/TECH.2011.08118

[33] Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasal, P., *et al*. (2001) Classification of Human Lung Carcinomas by MRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses. *PNAS*, **98**, 13790-13795. https://doi.org/10.1073/pnas.191502998