arXiv:2301.02283v1 [stat.ME] 5 Jan 2023

Scre基于非参数

贝叶斯测试

Naveed Merchant 和 Jeffrey D. Hart

2023年1月

摘要： 特征或变量选择是大型数据集固有的问题。尽管人们已提出许多方法来解决这一问题，但其中一些方法在数据集中的预测变量数量增加时，扩展性可能较差。筛选方法通过一次检查一个预测变量，使其随预测变量数量线性扩展；在进行进一步分析或变量选择之前，该方法可用于减少需要考虑的变量数量。对于分类，有多种技术可供使用。其中包括基于参数的筛选测试，例如 t-检验或基于 SIS 的筛选，以及基于非参数的筛选测试，例如基于柯尔莫哥洛夫距离的筛选 [Mai 和 Zou, 2012], 以及 MV-SIS [Cui 等, 2015]。我们提出了一种使用贝叶斯动机检验进行变量筛选的方法，将其与基于 SIS 的筛选进行比较，并通过模拟数据和真实数据上的示例应用来展示该方法。结果表明，我们的筛选方法能够提高分类率。即使将该方法与 DART 等旨在选择变量稀疏子集的分类器结合使用，也能取得这一效果。最后，我们提出了一种基于核密度估计的分类器，在某些情况下，相较于 DART，该分类器能够显著提高分类率。

Ke关键词：独立筛选、变量选择、分类、贝叶斯因子

1 引言

分类是指根据一组预测变量，预测某个观测值的类别标签。实现这一目标的技术种类繁多，包括支持向量机 [Boser 等, 1992], 基于树的方法 [Brieman 等, 1984], 贝叶斯树 [Linero 和 Yang, 2018] 以及梯度提升树 [Friedman, 2002]。

我们经常会遇到包含许多特征的数据集，但其中很少有特征全部都很重要。快速选取这些特征的一个子集是人们所期望的任务，但对于大型数据集而言可能颇为棘手。移除不重要变量后，前面提到的一些分类方法的性能可能会得到显著提升。特征选择并不是一个新领域，可以分为

三类：筛选或基于过滤的方法、包装方法和嵌入式方法[Tang 等人， 2014]。筛选方法逐一考察每个变量，以判断其是否提供有用信息，因此其计算规模随预测变量数量呈线性增长。然而，单独考察每个变量可能会导致变量联合行为的信息丢失，或导致选择共线变量 [Tang 等人， 2014]。包装方法拟合不同的模型，然后根据某项标准对每个模型进行评估。按照该标准表现最佳的模型将被选中。虽然这种方法往往能够选出较好的变量子集，但拟合每个模型可能会非常耗时，尤其是在数据集非常大时 [Tang 等人， 2014]。嵌入式方法生成一个内置某种稀疏性的模型，同时产生一组有用变量以及基于这些变量构建的模型。不同嵌入式方法的速度会因获取稀疏性所采用的策略而异，但它们通常比筛选方法更慢 [Tang 等人， 2014]。

本文将重点讨论筛选方法。先采用筛选方法，再采用嵌入式方法（如 LASSO），是常见的策略。Fan [Fan 和 Lv， 2008]采用了这一策略，并同时缩短了 LASSO 的运行时间，提高了整体模型的准确率。此后，出现了多种过滤方法。尤其是在分类领域，最大边际似然筛选 [Fan 等人， 2010], MV-SIS [Cui 等人， 2015], 以及柯尔莫哥洛夫距离筛选 [Mai 和 Zou， 2012] 都是已发表的筛选检验方法。这些方法大多应用于线性判别分析，并表明在筛选后将其应用于数据集能够带来改进。本文针对类别数已知为二的情形提出了一种新的筛选方法，并说明在不满足线性判别分析简单模型假设的情况下，该方法能够提高分类准确率。我们的筛选方法通过双样本贝叶斯检验识别有信息量的特征，该检验用于判断两个数据集是否具有相同的分布[Merchant and Hart， 2022]。我们的目标之一是表明，包括 BART 在内的分类方法 [Linero 和 Yang， 2018], DART[Linero 和 Yang， 2018] 以及支持向量机 [Boser 等人， 1992], 在采用我们的筛选程序后也能得到改进。

2 方法论

我们的筛选方法基于为每个特征单独计算一个统计量。我们将使用对应于两个类别的两个分布的核密度估计。其思路与柯尔莫哥洛夫距离筛选类似：如果某个特征的两个类别具有不同分布的可能性较大，我们就保留该特征。我们定义了一个检验统计量来作出这一判断。

假设我们观察到数据 $\boldsymbol{X}$，这是一个 $(m+n) \times p$ 矩阵，其第 i 行（X_{i1}，……， X_{ip}）包含某个研究对象的 p 个变量的取值。向量 $\boldsymbol{Y}$ 的第 i 个元素 Y_i 为 0 或 1，表示

第 i 个研究对象所属的类别，$i = 1, \ldots\ldots, m+n$。目前，假设 $Y_1, \ldots\ldots, Y_n$ 为 0，且 $Y_{n+1}, \ldots\ldots, Y_{n+m}$ 为 1。我们有 $n+m$ 个数据向量和 p 个特征。考虑特征 j 的数据 $\hat{\boldsymbol{X}}_j = (X_{1j}, \ldots\ldots, X_{(m+n)j})$，并定义 $h_i(\cdot|\boldsymbol{X}_j, b)$ 为带宽为 b、使用 $\boldsymbol{X}_j$ 中除第 i 个研究对象的数据之外所有数据的核密度估计：

$$\hat{h}_i(x|\boldsymbol{X}_j, b) = \frac{1}{nb} \sum_{\substack{r=1 \\ r \neq i}}^{n+m} K\left(\frac{x - X_{rj}}{b}\right).$$

我们还根据由满足 $Y = 0$ and$Y = 1$ 的观测值组成的数据集计算核估计。这些核估计为

$$\hat{f}_i(\cdot|\boldsymbol{X}_j, b) = \frac{1}{nb} \sum_{\substack{r=1 \\ r \neq i}}^{n} K\left(\frac{x - X_{rj}}{b}\right)$$

and

$$\hat{g}_i(\cdot|\boldsymbol{X}_j, b) = \frac{1}{nb} \sum_{\substack{r=n+1 \\ r \neq i}}^{n+m} K\left(\frac{x - X_{rj}}{b}\right).$$

然后，我们定义检验统计量 ALB_j、$j = 1$、……、p：

$$\begin{aligned}(m+n)ALB_j &= \sum_{i=n+1}^{m} \log\left(\frac{\hat{g}_i(X_{ij}|\boldsymbol{X}_j, b)}{\hat{h}_i(X_{ij}|\boldsymbol{X}_j, b)}\right) + \sum_{i=1}^{n} \log\left(\frac{\hat{f}_i(X_{ij}|\boldsymbol{X}_j, b)}{\hat{h}_i(X_{ij}|\boldsymbol{X}_j, b)}\right) \\ &\stackrel{\text{def}}{=} \sum_{i=n+1}^{m} \log(B_{1ji}) + \sum_{i=1}^{n} \log(B_{2ji}).\end{aligned}$$

记号 ALB 表示 平均对数贝叶斯因子，因为正如 [Merchant and Hart 所论证的，2022],统计量 B_{1jr} 和 B_{2js} 分别是基于数据“集” X_{rj} 和 X_{sj}的贝叶斯因子。我们参阅 [Merchant and Hart, 2022] 以深入了解此类平均对数贝叶斯因子。

ALB_j 的另一种解释基于 Kullback-Leibler 散度；对于密度 $h_1$and h_2，该散度为

$$KL(h_1, h_2) = \int_{-\infty}^{\infty} h_1(x) \log\left(\frac{h_1(x)}{h_2(x)}\right) dx.$$

定义 $f_{j,\ \text{mix}} = (nf_j + mg_j)/(n+m)$，其中 f_j 和 g_j 分别是类别 0 和 1 中特征 j 的密度。那么，统计量 ALB_j 是以下量的近似无偏估计量：

$$\left(\frac{n}{m+n}\right) KL\left(f_j, f_{j,\text{mix}}\right) + \left(\frac{m}{m+n}\right) KL\left(g_j, f_{j,\text{mix}}\right).$$

在这种情况下，当 m 和 n 趋于 ∞时，ALB_j 依概率收敛于 0。可以使用基于置换的程序评估ALB_j 的零分布，以确定两组观测值是否来自共同分布。

[Merchant and Hart, 2022] 建议选择带宽 b 时采用留一交叉验证，这一

种做法默认原假设成立。尽管这种策略有其潜力，但我们认为，针对每个变量分别采用这种方法的计算成本过高。我们只需执行此操作 p 次，这意味着计算量随变量长度线性增长，但这会使计算量随 n 呈二次增长，对于大型数据集而言难以承受。因此，我们选择将正态插件带宽与重尾 Hall 核结合使用，其形式为：

$$K_0(z) = \frac{1}{\sqrt{8\pi e}\,\Phi(1)} \exp\left[-\frac{1}{2}(\log(1+|z|))^2\right].$$

模拟结果表明，plug-in 中的常数精确到小数点后三位为 0.162，从而得到变量 j的以下 plug-in 规则：

$$b_{\text{plug-in},j}(\boldsymbol{X}_j) = 0.162(m+n)^{-1/5} s_j,$$

其中， s_j 是基础（零假设）标准差的估计值。一种可能的方法是将 s_j 取为 $\boldsymbol{X}_j$的样本标准差，但我们更倾向于采用更稳健的选择 $s_{Rj} = IQR(\boldsymbol{X}_j)/1.35$。

假设我们已经计算出 $ALB_1, ALB_2, ALB_3, \ldots,\ ALB_p$。接下来仍需为 ALBs选择一个截断值，以便选出所有 ALB 大于该截断值的变量。以下是几种可能的方法。

A1. 将截断值设为 $ALB_1, ALB_2, ALB_3, \ldots,\ ALB_p$的某个百分位数。这与 SIS 和 SIRS 的一些作者所提出的方法一致。假设我们预计数据集中有 d 个特征较为重要，那么可以将截断值设为第 $100(1-d/p)$百分位数。显然，最大的 d 个检验统计量最有可能具有显著性。这种方法的问题在于如何选择 d的适当值。SIS 的作者 [Fan 和 Lv， 2008] 以及 SIRS 的作者 [Zhu 等人， 2011] 认为，保守的选择是将 d 设为 n 或 $n\log(n)$，不过这些选择似乎有些武断。

A2. 一种经验性但计算上颇具挑战的方法是使用两个训练集和一种分类方法来处理这一问题。我们可以选择一个截断值，使分类方法在其中一个训练集上训练后应用于另一个训练集时的错误率最小。为使该过程的计算复杂度相对于 p保持线性，我们建议将候选截断值的数量限制在较小范围内，例如不超过十个。

A3. 随机选择一个协变量，例如 $\boldsymbol{X}_j$。对于该协变量，对标签进行置换，并计算检验统计量，记为 ALB_1^*。使用相同的特征 $\boldsymbol{X}_j$，重复此过程 d 次，得到 ALB_1^*、.....、 ALB_d^*。我们再不放回地随机选择 $B-1$ 个协变量，并对每个协变量重复前面的步骤，最终得到总共 Bd 个 ALB 值。完成后，我们将这些 Bd 值的一个百分位数选作截断值。此方法也能近似

ALB 的零条件分布（针对随机选择的特征），但它可能有一个优势：与 A3 相比，需要计算的统计量更少。

A4. 贝叶斯因子的解释 ALB_j 意味着，当 $ALB_j < 0$时，变量 j 绝不应通过筛选，这表明我们应将 0 用作截断值。虽然这个选择看起来可能较为宽松，[Merchant and Hart 2022] 证明，采用 ALB 0 作为截断值的测试，其第一类错误概率会随着 $m+n$ 趋于 ∞而趋于 0。

值得注意的是，特定的筛选方法与不同分类器搭配时会产生不同效果，尽管这或许在意料之中。我们建议使用能够利用位置差异以外的分布差异的分类器。一种常用的分类器是线性判别分析（LDA）；这里指的是假设两个类别具有相同协方差矩阵的版本。由于类别间尺度差异而被认定为重要的特征，通常对 LDA 毫无用处。

在 A1–A4 这些方法中，计算成本最低的是 A1 和 A4。使用这些截断值还有一个优点：当 $m+n$ 趋于 ∞时，测试效能趋于 1，因为 ALB_j 在 $f_j \equiv g_j$ [Merchant and Hart 2022]下依概率收敛于 0。如果能够明确预期相关变量的数量，或者另一个分类器正常运行所需的变量数有一个临界值，则建议采用方法 A1。

最后，我们想指出，每个 ALB 都有一个有限的上界，因为很容易证明

$$ALB_j \leq \log(2) \cdot \max\left(\frac{m}{(m-1)}, \frac{n}{(n-1)}\right),$$

这意味着，只要 m 和 n 不太小， ALB_j 就基本有界于 log(2)。

3 一致性结果

我们首先假设每个变量都满足 [Merchant and Hart, 2022] 中的条件 A1-A5 .

我们还假设，两个类别的样本数 m 和 n趋于无穷大，而变量数 p固定。假设当某个变量能够单独提供信息时，该变量属于类别 D；这意味着该变量在一个类别中的分布与在另一个类别中的分布不同。最后，我们假设这些变量相互独立。

定理1。 如果上述假设成立，那么

$$\lim_{n,m\to\infty} P(\max_{i\in D^c} ALB_i < \min_{j\in D} ALB_j) \to 1. \tag{1}$$

*S*假设我们采用 *A4* 中的截断值（在我们的第*3*节中）。那么我们还可以得到以下结果：

$$\lim_{n,m\to\infty} P(\min_{j\in D} ALB_j > 0 \cap \max_{j\in D^c} ALB_j < 0) \to 1. \tag{2}$$

证明。结果（2）蕴含（1），因此我们只需证明前者。利用以下事实 P（$A \cap B$） $\geq 1 - P$（A^c） $- P$（B^c），只需证明 P（最小值$_{j\in D}\, ALB_j > 0$）和 P（最大值$_{j\in D^c}\, ALB_j < 0$）都趋于1。

We 只考虑 P(最大值$_{j\in D^c}\, ALB_j < 0$)，因为另一个概率的证明类似。我们有 e

$$\begin{aligned} P(\max_{j\in D^c} ALB_j < 0) &= P\left(\bigcap_{j\in D^c} \{ALB_j < 0\}\right) \\ &= \prod_{j\in D^c} P(ALB_j < 0) \\ &\geq \left(\min_{j\in D^c} P(ALB_j < 0)\right)^N, \end{aligned} \tag{3}$$

其中 N 是 D^c中的元素个数。对于每个 $j \in D^c$， [Merchant and Hart, 2022] 证明P（$ALB_j < 0$） $\to 1$ 当 m 且 n 趋于 ∞时。由于 N 是有限的，这意味着（3）右端的量趋于1，结论由此得出。 ■

显然，如果 p 以足够慢的速度趋于 ∞，那么定理1仍然成立。然而，要确定 p 相对于 m 和 n 的增长速度，需要比 [Merchant and Hart 2022], 所提供的更强结果，我们不再进一步探讨这一方向。

现在，我们展示 $m = n$、 p 和 r取不同值时的模拟结果，其中 r 表示重要变量所占的比例，即类别分布存在差异的变量所占的比例。我们按以下方式为一个二分类问题生成 500 个变量。如果变量是重要变量，那么在一个类别中，该变量取自自由度为 4 的 t分布，在另一个类别中则取自两个正态分布的混合，其中混合参数为 1/2，两个正态分布的标准差均为 1，均值分别为 -2.5 和 2.5。反之，如果变量不重要，则该变量取自标准正态分布。我们将这种情形下生成变量的方法称为“形状差异”。最后，通过一次成功概率为 $r = 1/2$的二项试验，确定每个变量是否重要。图1、2和3展示了 ALB的累积分布函数如何随 m 和 n而变化。

随着样本量 m 和 n 增加，重要变量的 ALBs逐渐增大。即使观测值总数仅为变量总数的百分之八，我们也能达到这样的性质：不重要变量中最大的 ALB 小于重要变量中最小的ALB 。黑色曲线显示的是通过对每个变量置换标签三次，并且每次计算 ALB，从而计算得到的 ALBs的累积分布函数。截断值为 0 并不大于任何 m 下不重要变量的最大值，

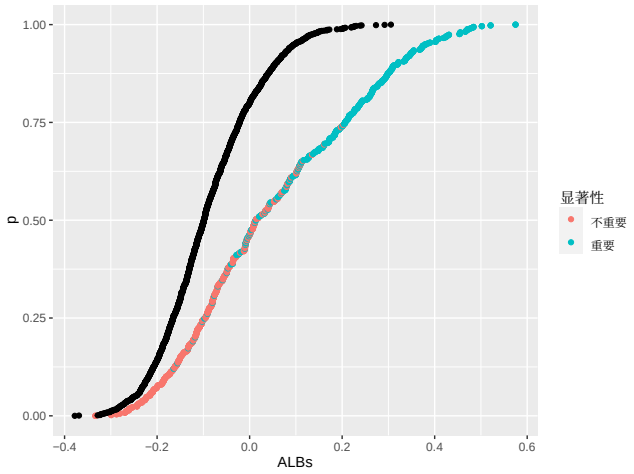


图1： 训练集大小为*10*且重要变量存在形状差异时 *ALB* 累积分布函数的比较。黑色曲线表示由 ALB 生成的累积分布函数，其中数据的类别经过了置换。

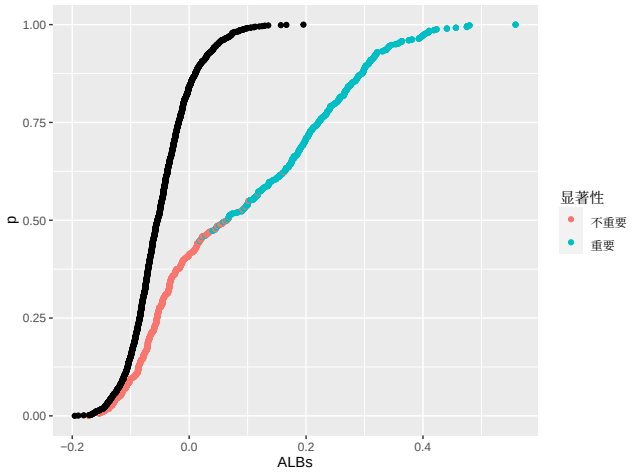


图2： 训练集大小为*20*且重要变量存在形状差异时 *ALB* 累积分布函数的比较。黑色曲线表示由 ALB 生成的累积分布函数，其中数据的类别经过了置换。

但仍有助于剔除大量不重要变量。另一方面，使用较大的置换变量百分位数可能会导致几乎所有不重要变量都被剔除；在最大样本量下，将截断值设为置换 ALBs的最大值，确实几乎可以完美地区分重要变量和不重要变量。

4 分类方法讨论

理想情况下，应选择彼此配合效果最佳的筛选方法和分类器。以下关系很好地体现了这一原则：逻辑回归、支持向量机和线性判别分析等分类方法与基于 t-检验的筛选方法之间的关系——

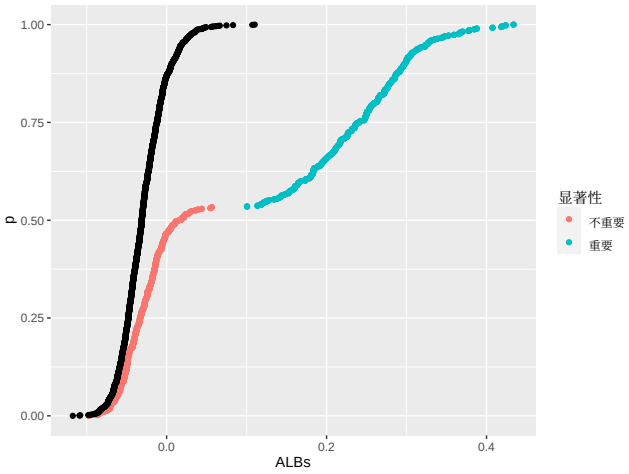


图 3： 训练集大小均为 *40* 且重要变量存在形状差异时 *ALB* 累积分布函数的比较。黑色曲线表示以下数据所生成的累积分布函数： ALB，其中类别经过了置换。

方法。判别分析、不使用核技巧的支持向量机以及逻辑回归，都是为利用类别之间的位置差异而设计的。因此，先采用 t-检验进行筛选是很自然的做法；当然，这种检验旨在检测均值差异。另一方面，使用核技巧的支持向量机会在进行变换后，构造出能够最佳分隔两个类别的超平面，因此能够有效处理多种分布差异。为了发挥这一优势，最好使用能够检测非位置差异的筛选方法。总之，当线性判别分析或逻辑回归被认为是合适的分类器时，采用 t-检验筛选是自然的选择；但若需要使用带核技巧的支持向量机，这种方法未必合适。线性支持向量机是指通过一个平面分隔两组来进行分类的方法。

在图4 和5 中，两个（重要）变量均按照形状差异生成。两个类别分布之一的双峰性使得难以用一个平面将这些类别分开。图 4 展示了仅使用两个相关变量进行分类时支持向量机的性能，而图 5 展示了在相同情况下应用核技巧检测非位置差异所带来的改进。

我们的方法旨在通过考虑位置类型以外的差异，优于 t-测试筛选。当然，这种性能并非没有代价：我们在检测均值差异时会损失一些检验效能。若要让我们的方法与分类器配合得更好，分类器必须能够区分表现出非位置差异的类别。例如，如果一组变量所属的类别仅在尺度上有所差异，那么对于不使用核技巧的支持向量机和 LDA 而言，这组变量并无用处，因为不存在能够恰当地分隔这些类别的超平面。采用核技巧，或通过考虑交互作用和平方

项来增加变量数，可以绕开这个问题。但增加变量并非理想做法，因为我们的方法论旨在降低计算复杂度。

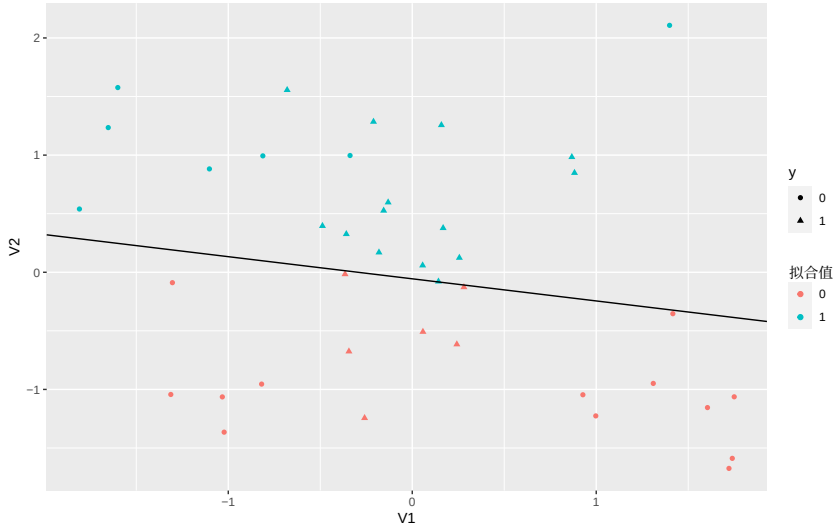


图4：使用线性支持向量机，在“形状差异”情形下对两个重要变量进行分类的预测准确率。颜色表示支持向量机所做出的预测。直线表示线性支持向量机生成的判别边界，用于区分类别。三角形和圆形表示观测值所属的类别。

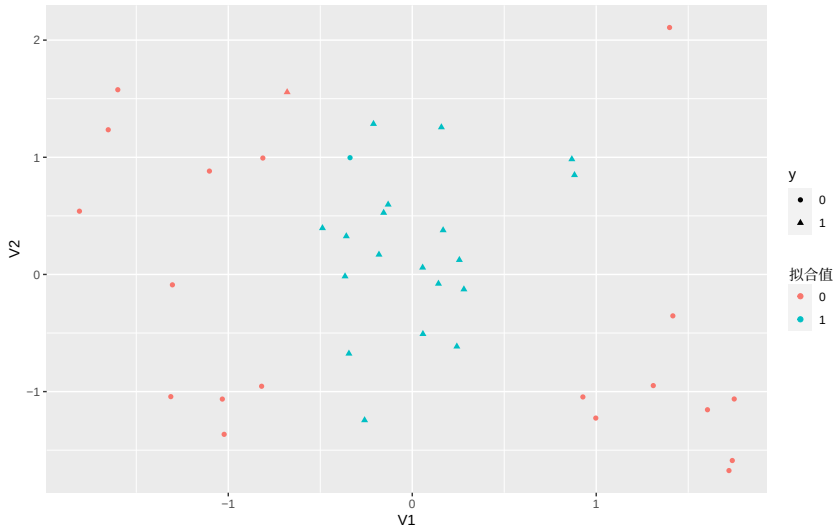
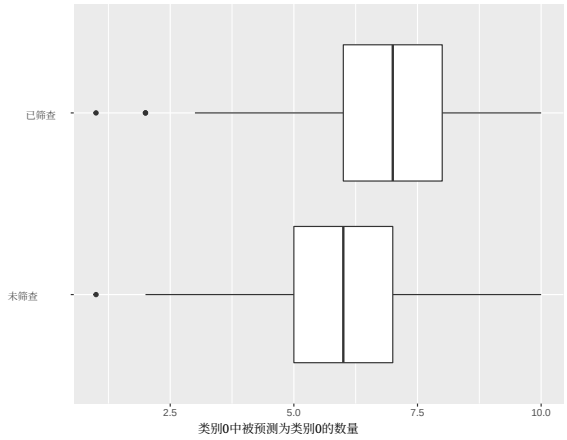
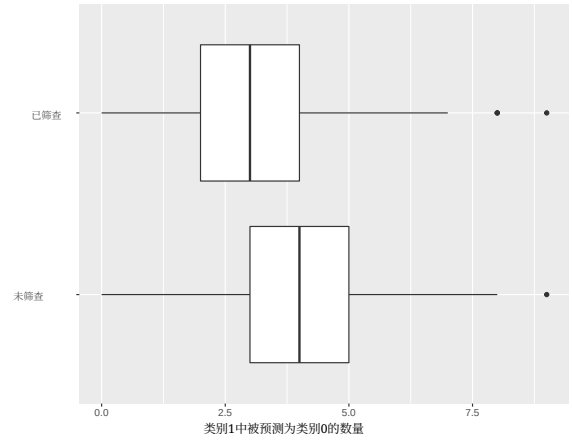


图5：使用采用核技巧的支持向量机，在“形状差异”情形下对两个重要变量进行分类的预测准确率。颜色表示应用核技巧后支持向量机所做出的预测。三角形和圆形表示观测值所属的类别。在这种情况下，分类效果要好得多，因为核技巧使分类器能够捕捉位置偏移以外的差异。

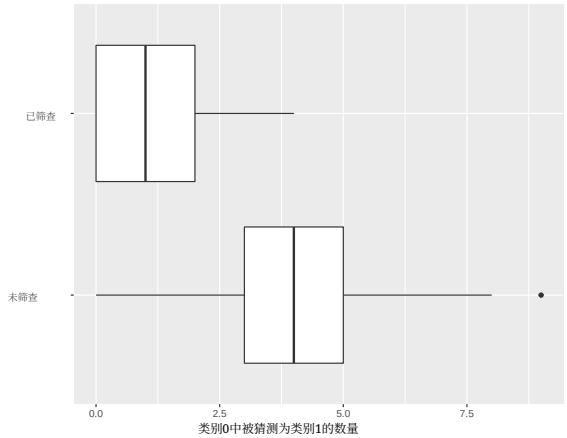
最后，我们想补充说明，尽管采用核技巧的支持向量机是一种出色的分类方法，能够揭示变量之间多种不同类型的差异，但噪声变量的存在会严重降低其性能。为说明这一点，我们使用与形状差异相同设置下的数据。我们构建了训练集和测试集，两者均包含



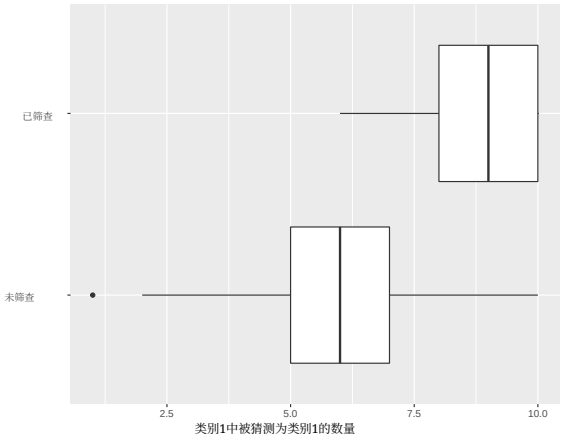
(a) 真正例的箱线图



(b) 假正例的箱线图



(c) 假负例的箱线图



(d) 真负例的箱线图

图6：展示数据由某个模型生成时支持向量机准确率的箱线图，其中10% 的变量具有重要性且存在形状差异。为了说明这些方法的准确率，我们进行如下操作。首先，假设正例对应于观测值属于类别 1，负例对应于观测值属于另一类别。然后，为了展示支持向量机的准确率，我们绘制箱线图，显示“真正例”“真负例”“假正例”和“假负例”的出现次数。

每个类别的 10 个观测值。使用了 500 个变量，其中平均只有 10% 是重要变量。所有不重要变量的数据均服从标准正态分布。如果某个变量是重要变量，那么它在一个类别中的分布是两个正态分布的双峰混合分布，而在另一个类别中则服从 t-分布，自由度为 4。混合分布中的两个正态分布的标准差均为 1，均值分别为 -2.5 和 2.5。我们使用所有观测值训练了一个采用径向基核的支持向量机，并训练了另一个采用径向基核的支持向量机，但仅使用 ALB 值大于解释性截断值 0 的变量。我们重复这一过程 100 次，并在图 6 中报告其准确率。总体而言，分类准确率大幅提高；筛选后，假阴性很少出现，假正例的数量也有所减少。

ALB筛选方法不必最终决定应纳入哪些变量。两个变量可能各自都很重要，但高度相关，因此可能会同时被选中，尽管这对某些分类器而言

可能并不理想。筛选可以仅仅作为一种前置步骤，简化分类器的工作，而分类器会进一步进行变量选择。即使是能够同时进行变量选择和建模的方法，也能受益于通过筛选大幅减少变量数。在 SIRS [Zhu 等人， 2011], SIS [Fan 和 Lv， 2008], 以及我们的案例中，情况也是如此。

5 与 BART 的交互及定制分类方法

我们建议采用能够利用类别间存在非位置差异的特征的分类方法。进一步进行变量选择或能够处理稀疏数据集的方法，也能与我们的筛选方法配合得很好。BART 和 DART 对参数假设较少，并且能够从数据中捕捉多种特征。随着预测变量数量增加，BART 会遇到问题；DART 则被提出作为解决这一问题的方法 [Linero 和 Yang， 2018]。DART 虽然能够处理许多预测变量无关的情况，但也要付出代价。与 BART 相比，DART 的马尔可夫链混合时间更长；此外，鼓励稀疏性的先验分布可能导致 DART 在用于估计它的 MCMC 过程运行时陷入某个后验众数 [Hill 等人， 2020]。虽然我们无法直接缓解这些问题，但减少变量数有助于加快 MCMC 过程。与 DART 相比，我们的筛选方法能够更快地减少变量数。DART 能够抵御相关的干扰变量，因此可以剔除那些通过 ALB 筛选、但因共线性而无关紧要的变量。我们通过模拟表明，在 BART 或 DART 之前使用我们的筛选方法，可以降低误分类率并提升计算速度。

我们在与图 2相同的情境下生成数据，但相关变量约占变量总数的 10%。若某个变量无关，则该变量在两个类别中的分布均为标准正态分布。为了评估分类器的性能，我们计算了兰德指数，即分类器作出正确判断的比例。我们分别计算 BART 和 DART 方法应用于所有变量时的兰德指数，以及这两种方法应用于通过 ALB 筛选的变量时的兰德指数。我们考虑了训练集大小从 5 到 20 不等的情况。每次模拟的测试集大小与训练集大小相同。对于每个样本量，我们重复这一过程 100 次。

图 7 和9 展示了在这些设置下，单独使用 BART 和 DART 的准确率；图8 和 10 展示了预先筛选出重要变量后，各方法的表现。随着训练集大小逐渐增加，两种方法的兰德指数均有显著提升。更值得注意的是，运行这两种流程所需的时间减少了。图 11 展示了筛选前运行 BART 方法所需的时间，图12 展示了筛选后该方法所需的时间。筛选平均至少可减少 10 秒的计算

时间，同时提高平均准确率。这一结果很有意思，因为这些方法本身，尤其是 DART，往往能够抵御无关变量的影响。不过，图中结果表明，要实现这种稳健性，需要更大的样本量。

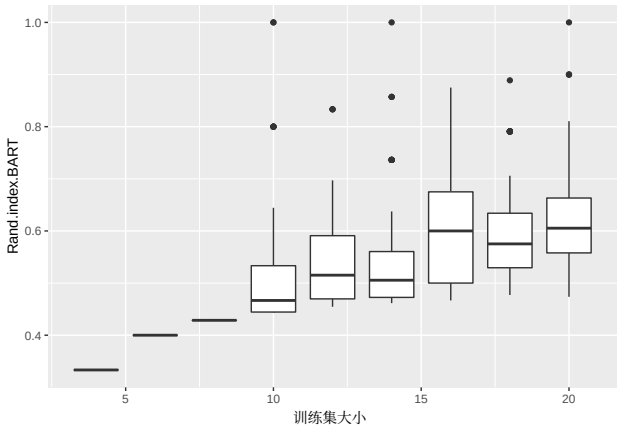


图7：形状差异情形下 *BART* 兰德指数的箱线图。我们改变 m with$m = n$，图中的训练集大小表示 $m + n$。对于每个样本量，我们将每次模拟重复 100 次。大约 10% 的变量是相关变量。

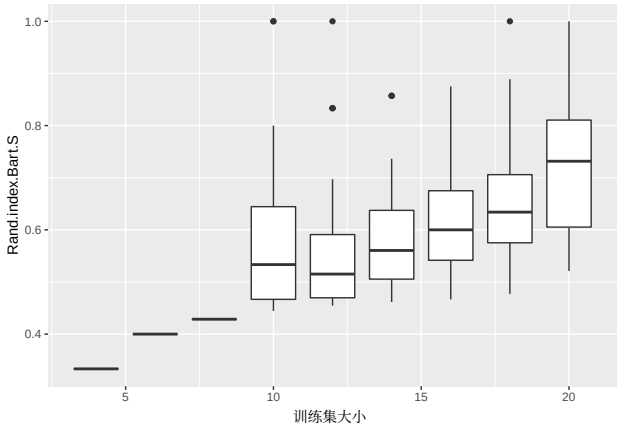


图8：通过筛选改进 *BART* 后，*BART* 兰德指数的箱线图。设置与图7相同。两个图的区别在于，我们在应用 BART 之前使用 ALB 程序筛选了变量。我们选择变量，使得所有生成的 ALB 均大于 0 的可解释截断值。当累积的数据足够多时，分类方法的检验效能会变得很高。解释性阈值往往较为保守；如果改用基于置换的截断值，该方法的检验效能可能会更高。

5.1 一个简单的贝叶斯分类器

假设我们的目标只是利用变量之间的差异，而不论差异的类型，并且假定变量之间相互独立。我们可以按以下方式构建一个简单的贝叶斯分类方法。对于每个变量，我们计算两个核密度估计，每个类别各一个。对于每个变量 i，令 f_i 和 $\hat{}g_i$ 分别为使用类别 1 和类别 2 中所有变量 i 数据计算得到的核密度估计。变量来自某一

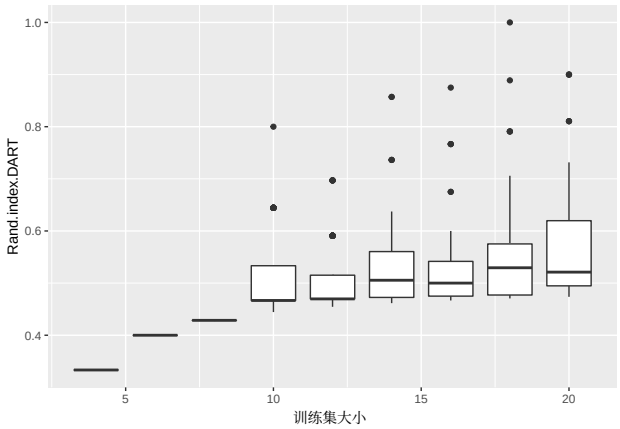


图9：*DART* 兰德指数的箱线图。此处的设置与图7相同，不同之处在于我们使用 DART 而非 BART，因为 DART 能够自动进行变量选择。

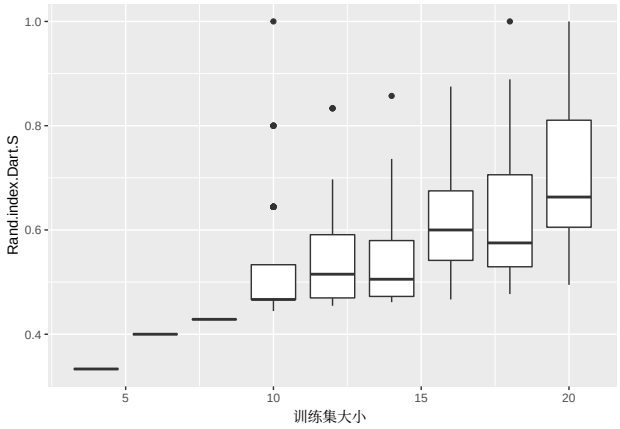


图10：经筛选改进后的 *DART* 兰德指数箱线图。此处的设置与图7相同。不同之处在于，我们在应用 DART 之前使用 ALB 程序筛选变量。我们选择变量，使所有生成的 ALB 均大于 0。尽管 DART 完全能够自动进行变量选择，筛选变量后分类方法的检验效能仍有所提高。

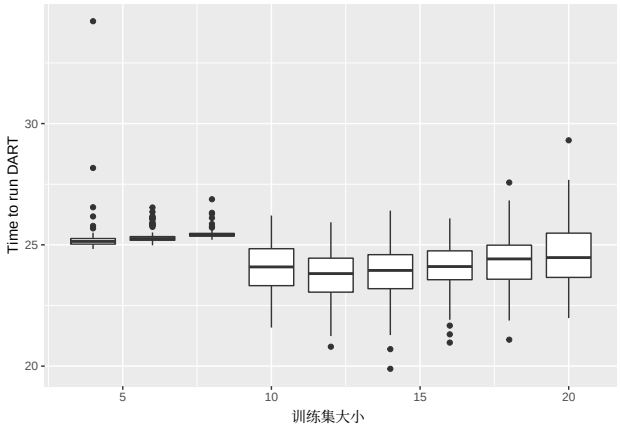


图11：运行 *DART* 所需时间的箱线图。模拟采用与图7相同的设置。

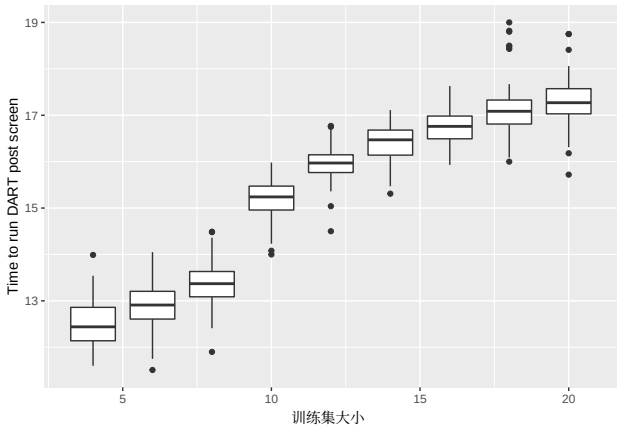


图12： *DART* 在筛选后运行所需时间的箱线图。 设置与图 7中的相同。此处的区别在于，应用 DART 之前，我们先使用 ALB 程序筛选变量。筛选过程本身耗时远不到半秒；由于移除了大量无关变量，DART 的运行时间大幅缩短。

类别被假设为与该类别的观测值数量成正比。设 $\boldsymbol{x} = (x_1, ..., x_p)$ 为一个待分类的观测值。若已知相应的密度，则 $\boldsymbol{x}$ 来自类别1的条件概率为

$$p(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = \frac{\frac{n}{m+n}\prod_{i\in D} f_i(x_i)}{\frac{n}{m+n}\prod_{i\in D} f_i(x_i) + \frac{m}{m+n}\prod_{i\in D} g_i(x_i)}, \tag{4}$$

其中， D 是满足 $f_i \neq g_i$ 的索引集合 i 。当然，密度和 D 是未知的，但 $\frown p(\boldsymbol{x})$ 可以用核密度估计进行估算，而 D 可以替换为 D，即相应变量通过筛选后所对应的索引集合：

$$\hat{p}(\boldsymbol{x}) = \frac{\frac{n}{m+n}\prod_{i\in \widehat{D}} \hat{f}_i(x_i)}{\frac{n}{m+n}\prod_{i\in \widehat{D}} \hat{f}_i(x_i) + \frac{m}{m+n}\prod_{i\in \widehat{D}} \hat{g}_i(x_i)}. \tag{5}$$

基于（5）的分类器可以有效捕捉分布中的边际差异，但无法利用变量依赖结构中可能存在的差异。为了使用（5），如果观测值 $\boldsymbol{x}$ 满足 $\hat{}p$（$\boldsymbol{x}$）$> n/$（$m+n$），我们称其属于类别1，否则属于类别2。我们发现，在处理相互独立的变量，或多变量分布之间的差异主要由边际差异决定的情形时，该分类器具有很高的准确率。

为了说明基于（5）的分类器的有效性，我们在不同样本量下进行了与图1 所引用的模拟类似的模拟。我们将这一过程重复 5 次。对于每种样本量，我们生成一个大小相同且类别均衡的测试集，并计算兰德指数。图13展示了兰德指数与样本量之间的关系。即使在较小的 $m + n$ 值下，该分类方法的检验效能也会变得相当高。由于解释性阈值往往较为保守，如果改用基于置换的阈值，该方法的检验效能可能会更高。每组训练样本量仅为 9 时，分类器所做的

完美预测有超过 75% 的概率。

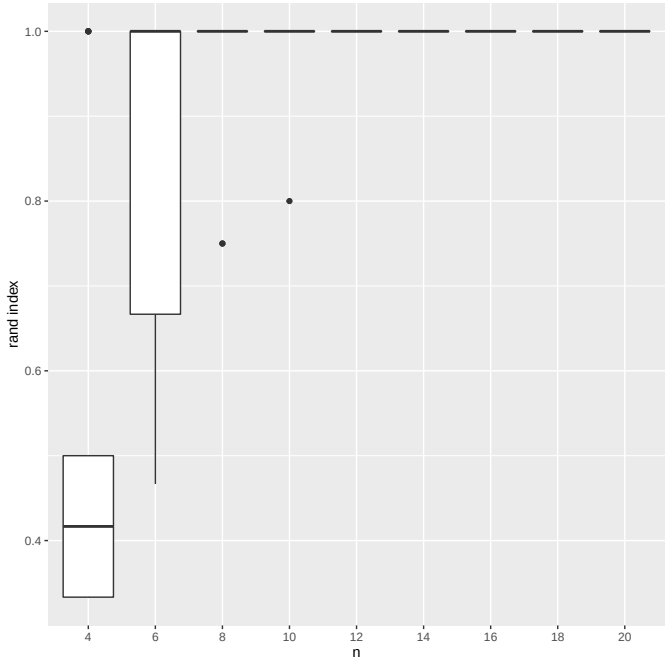


图13：贝叶斯分类器兰德指数的箱线图。该分类器根据（5）定义。我们在与图2相同的情境下生成数据，但改变 m 和 n，使得 $m = n$，训练集大小表示 $m + n$。我们选择变量，使生成的所有 ALB 均大于具有可解释性的 0 阈值。

6 模拟数据集上的应用

我们想将我们的方法与 t-检验筛选进行比较，也比较不同 ALB 阈值选择的表现。至少有两种方法可以进行比较。一种是比较两种方法筛选后保留下来的变量百分比；另一种是在筛选后应用分类方法，并观察哪种方法的分类率更高。我们在以下三种情形下进行这些步骤：

情形 1 – 位置差异。共有 600 个变量，其中重要变量对应的情形是类别之间存在均值差异。如果变量重要，则一个类别服从标准正态分布，另一个类别服从均值为 1、标准差为 1 的正态分布。我们通过生成 600 个相互独立的伯努利变量（每个变量的成功概率为 0.05），使约 5% 的变量为重要变量。在情形 2 和情形 3 中，也采用这种方法来确定重要变量。此外，在本情形及接下来的两种情形中，不重要变量均服从标准正态分布。在本情形中，我们将使用不带核技巧的支持向量机来比较表现。

C例 2 – 尺度差异。 共有 600 个变量，其中重要变量来自 m

类别之间存在方差差异的情形。如果变量很重要，则一个类别服从标准正态分布，另一个类别服从均值为 0、标准差为 3 的正态分布。我们设定约 20% 的变量为重要变量，所用分类器为采用核技巧的支持向量机。

案例 3 –形状差异。共有 600 个变量，其中重要变量来自类别分布形状不同的情形。如果变量重要，则一个类别服从自由度为 4 的标准 t分布，另一个类别服从由均值为 -2.5 和 2.5、标准差均为 1 的两个正态分布组成的双峰混合分布。约 10% 的变量是重要变量，所用分类器为采用核技巧的支持向量机。

对于这三种情形，均进行了筛选，并使用每个变量上 $m+n$ 个观测值构成的训练集建立分类器，其中 $m = n$。随后，使用该分类器预测 $m + n$ 个观测值，并计算所得兰德指数。在 t-检验筛选中，当其P-值小于 0。005 时，便选中相应变量。对于每个样本量，我们重复此过程 100 次。图14-16 分别展示了采用三种方式选择 ALB 阈值时各方法的表现：“固定第一类错误率”方法、ALB 的最大 $n + m$ 值，以及 0 的截断值。

这些模拟结果表明，ALB筛选能够有效检测位置差异（如案例 1 所示），但效果不及 t-检验筛选。在案例 1 中，采用 ALB 筛选的支持向量机性能优于不进行筛选，但不如采用 t-检验筛选。随着样本量增加，通过 ALB 筛选的变量比例稳步上升，但增速低于采用 t-检验筛选的情况。在案例 2 和 3 中，就分类准确率而言， t- 检验筛选并不比不进行筛选效果更好。至于重要变量的保留，随着样本量增加， t-检验筛选的效果并未改善，但 ALB 筛选有所改善。

7 GISETTE 数据应用

GISETTE 数据来自 $m = 3000$ 和 $n = 3000$ 手写图像，分别为数字 4 和 9。对于 6000 张图像中的每一张，我们测量了 $p = 5000$ 个变量，其中一些是无关探针，其余为像素强度。我们将使用不同的筛选方法选择不同的变量子集，并对这些数据进行分类。我们将以 DART 作为主要分类方法，并探究在不同筛选方法的辅助下，其表现如何。

数据集被随机分为两半。第一半作为训练数据。我们使用这些数据训练分类器，并计算了 t-统计量和 ALB 统计量。第二半作为验证集，我们利用这些数据计算兰德指数。

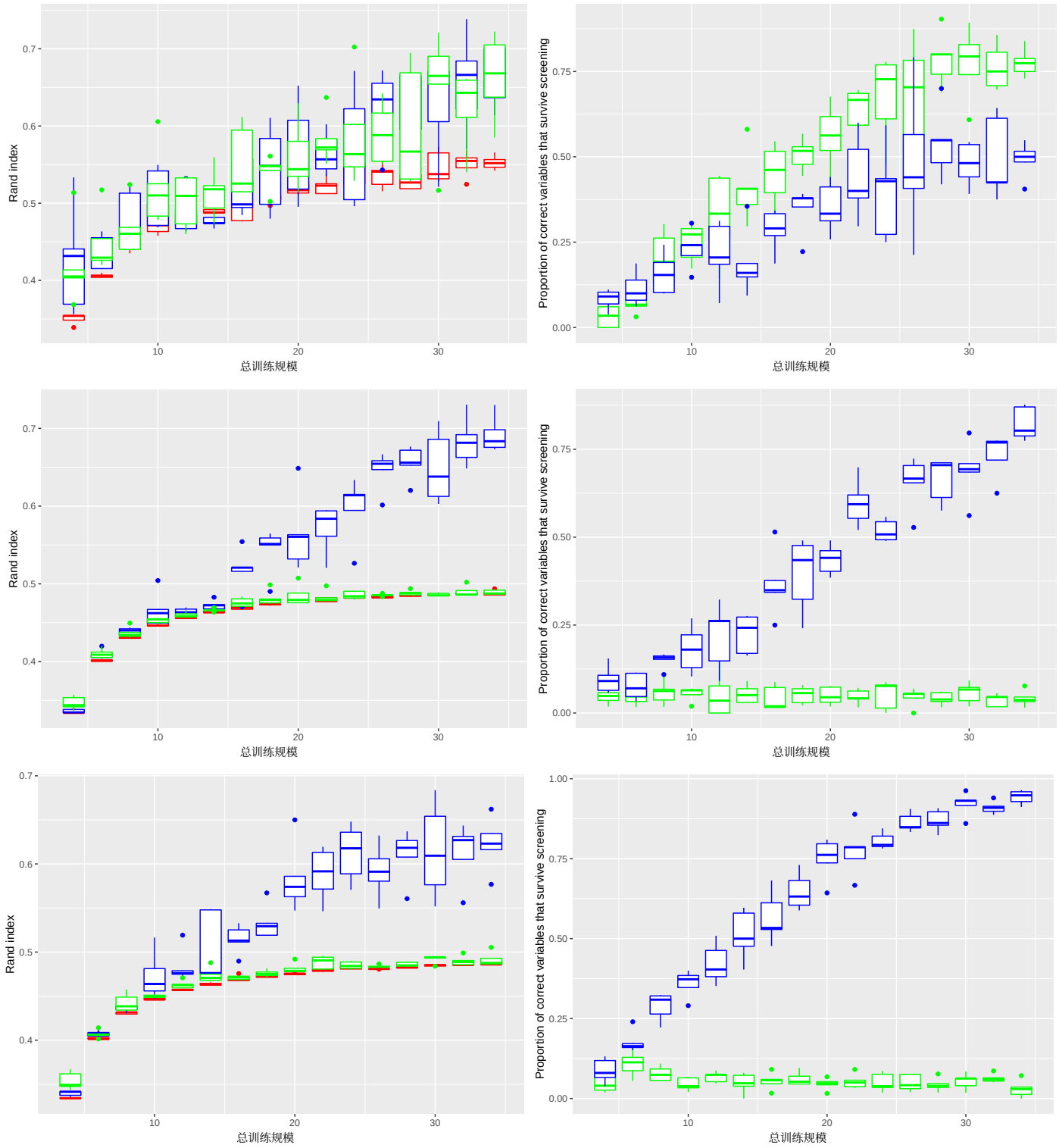


图14：基于 t-检验筛选和采用 $A3$ 的 ALB 筛选的模拟结果。我们通过置换数据标签并计算所得数据的 ALB，生成ALB^* 值。对于每个变量，我们都进行两次此操作。ALB阈值为 ALB^*的第95百分位数。选择 t统计量的截断值，使得 P值小于0。05。红色、绿色和蓝色箱线图分别表示不进行筛选、进行 t-检验筛选和进行 ALB 筛选的结果。第一、第二和第三行图分别对应情况1、2和3。

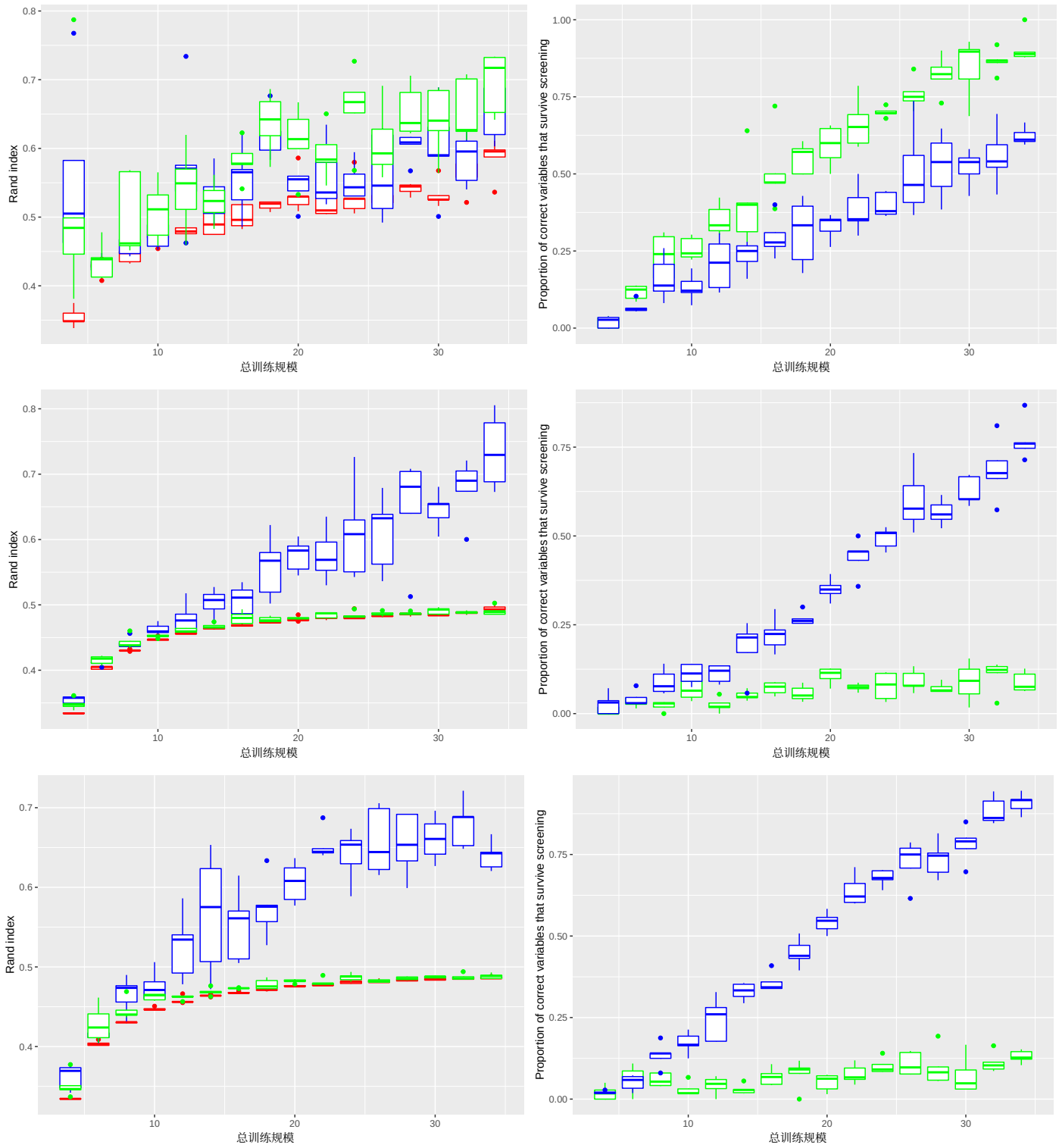


图15： 针对 t-测试筛选和 ALB筛选的模拟结果，其中使用了具有 $n+m$最大 ALB 值的变量。 箱线图的颜色含义与图 14 中相同。

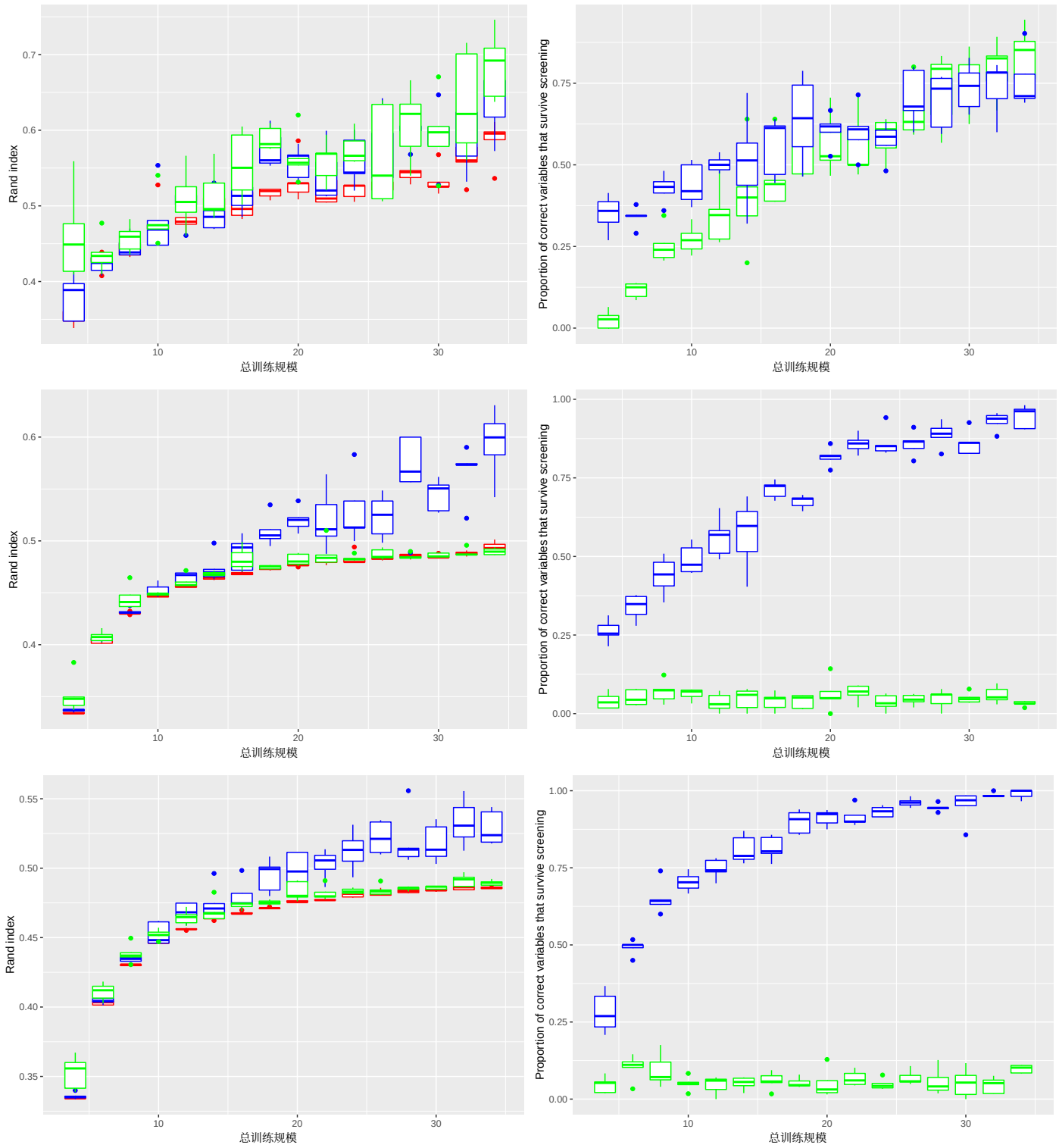


图 16： 的模拟结果-测试筛选和 ALB筛选，截断值为0。箱线图的颜色含义与图14中的相同。

兰德指数	筛选方法	所选变量数
0.947	$ALB > T^*_{0.005}$	1321
0.942	$ALB > 0$	1946
0.947	$P_t < 0.005$	1540
0.935	不进行筛选	4835

表 1：*GISETTE* 数据的分类和筛选结果。所有方法均使用了平衡训练集和测试集，其中训练集和测试集都包含 3000 个观测值。 P_t 和 $T^*_{0.005}$ 分别是一次 t-测试的 P值，以及置换 ALB的第99.5百分位数。

我们比较了五种筛选方法。我们采用 A3，设定 $B = 1000$ 和 $d = 1$，并选择满足以下条件的变量： ALB 大于 。 值的第 99.5 百分位数。为了将其与 t-检验筛选进行比较，我们选取 t-检验 P- 值小于 0.005 的变量。我们尝试了筛选方法 A1，并将其与选取相同数量变量的对应 t 检验方法进行比较。我们还考察了不进行变量筛选时该方法的表现。我们将这些方法的兰德指数与不进行变量筛选时的兰德指数进行比较。在采用任何方法之前，我们删除了所有 6000 个数据值都相同的变量。因此，完整数据集中只有 4835 个变量，而不是 5000 个。兰德指数汇总见表 1。

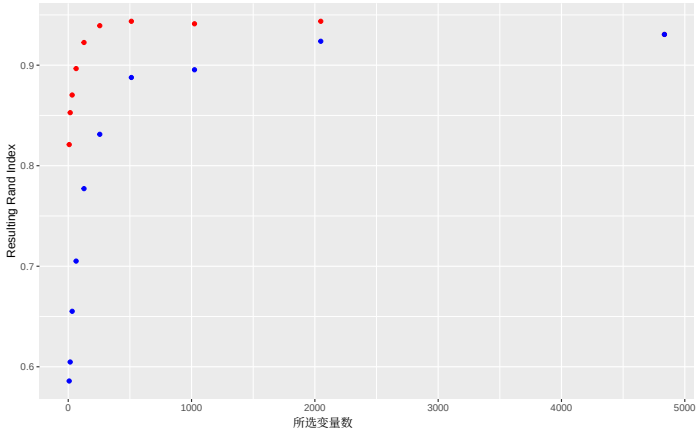


图 17：*GISETTE* 数据的兰德指数随所用变量数的变化。红色点对应 ALB 筛选，蓝色点对应 t 统计量筛选。兰德指数是针对 DART 分类器计算的，该分类器仅使用 2^j t 统计量最大或 ALB 的 2^j 值最大的变量，其中 $j = 3$，.....，11。

对于这些数据，基于 A4 和第 99.5 百分位数的 t-test 筛选与 ALB筛选效果相当。两种方法的区别在于，ALB 在保持相同准确率的同时，筛选出的变量大约少 200 个。可解释截断规则的效果最差，但仅使用四个“像素”测量就达到 75% 的准确率，这一结果非常有趣。我们认为，在这种情况下，大多数边际上重要的变量（或“像素”）是那些在数字 4 或 9 中某一个数字上着色、而在另一个数字上未着色的像素。这可以解释为位置差异，因为着色像素的强度大于很少被触及的像素的强度。我们还认为，这

也是 t-test 筛选在相同 I 类错误率下比 ALB筛选找到更多重要变量的原因。尽管这种情况下主要存在位置差异，但至少在 I 类错误率为 0.005 时，ALB 的兰德指数表现最终与 t-test 筛选相当。总体而言，我们认为，对于 n 和 p 都较大的数据集，依据 I 类错误率选择截断值或将 0 用作截断值，都是不错的策略。大量数据使我们能够采用相当小的显著性阈值，同时仍保持良好的检验效能。在这种情况下，选择 ALB 的 $n+m$ 值最大的变量会导致没有变量通过筛选，因此我们决定不再探究这一做法。由于 p 和 $n+m$的值很大，使用交叉验证程序选择截断值的成本很高。

为探究基于分位数选择截断值的影响，我们考察了当截断值对应于使用具有最大 k 统计量的变量时的兰德指数。我们考虑了按几何级数递增的 k 取值： $k=8$、 16、.....、 2048。结果见图 17。在使用的变量数相同的情况下，基于 ALB 的筛选比 t-检验筛选具有更大的兰德指数。显然，ALB 和 t-检验筛选选择的并非相同的变量，而 ALB 选择的变量更为有效。

8 白血病数据的应用

白血病数据集包含 72 名患者的观测值，其中 47 名患有一种类型的白血病，其余患者患有另一种类型。我们获得了每位患者的 7129 个变量的观测值，以构建一个分类器，帮助判断未来患者所患的白血病类型。这是一个 p 远大于 n 和 m的情形。我们将把各种筛选方法应用于这一问题，并结合 DART，通过兰德指数比较这些方法的准确率。

为公平起见，我们将数据集随机分成两半。前一半作为验证数据，后一半作为训练数据。然后再将训练数据对半分开，得到两个较小的训练集。这样做有两个原因。首先，我们希望评估训练集大小对各方法准确率的影响。两个较小训练集中的一个将用于构建分类器，每个分类器对应一种不同的筛选方法；随后，我们将使用全部训练数据构建另一组分类器。两组分类器都将用于预测验证集中的数据。将训练集对半分开的第二个原因是，这样便可以采用交叉验证方法来选择截断值。我们可以在其中一个较小的训练数据集上训练模型，并选择一个能在另一个训练数据集上取得最佳分类准确率的截断值。然后，我们可以在完整训练集上训练这个最佳模型，并将其应用于验证集。这是在方法论中采用策略 A2；由于 m 和 n 的规模较小，这种做法是可行的。

为开展分析，我们采取了以下步骤。我们实施了四种基于 ALB 的筛查方法

兰德指数	所用筛选方法	变量数
0.599	$n+m$ 最大的 ALBs	19
0.529	$ALB > T^*_{0.05}$	617
0.599	$n+m$ 最大的 t-统计量	19
0.529	$P_t < 0.05$	847
0.501	不进行筛选	7129

表 2：训练集为完整数据集的四分之一时，白血病数据的分类和筛选结果。所有结果均基于大小约为完整数据集一半的验证集。量 P_t 和 $T^*_{0.05}$ 分别是 t 检验的 P 值以及置换后的 ALB 的第95百分位数。所用的分类器为 DART。

兰德指数	筛选方法	变量数
0.742	$ALB > T_{\mathrm{CV}}$	12
0.742	$ALB > T_{\mathrm{CV}}$	12
0.786	$n+m$ 个最大的 ALB	38
0.572	$ALB > T^*_{0.05}$	1302
0.598	$n+m$ 个最大的 t 统计量	38
0.572	$P_t < .05$	1694
0.549	不进行筛选	7129

表3： 训练集为完整数据集一半时，白血病数据的分类和筛选结果。所有结果均基于一个验证集，其大小约为完整数据集的一半。表中的前两行对应于使用基于（5）的分类器，后续各行对应于使用 DART。数值 T_{CV} 是通过交叉验证选出的最佳截断值， P_t 和 $T^*_{0.05}$ 与表2中的相同。关于交叉验证的具体实施方式，请参见第8 节。

以及两种基于 t-检验的筛选方法。四种 ALB 方法分别为：A1，选取最大的 $n+m$ 个ALBs；A2，截断值为 0；A3，I 类错误率为 0.05； $d = 3$ 和$B = 7129$。

使用了两种 t-检验筛选方法：一种选取具有最大 $m+n$ t-统计量的变量，另一种选取其 t-检验 P-值小于 0.05 的变量。后一种 t-检验筛选方法使其可与采用显著性水平 0.05 的 A3 方法选择变量相比较。通过筛选确定相关变量后，使用基于这些变量的 DART，从验证集计算兰德指数。表 2 和 3 汇总了结果。

当训练集大小为 $m+n$的四分之一时，所有筛选方法的表现都相近。然而， ALB 采用 $A1$ 或 $A3$ 中的方式选择截断值的筛选方法，在样本量翻倍后表现显著提升，远胜于基于 t检验的筛选方法。虽然已有研究表明 DART 是一种有效的分类器，但根据我们的经验，当样本量较小时，它可能无法恢复分类问题的结构。因此，我们尝试使用基于（5）的贝叶斯分类器，并采用基于交叉验证的方法选择截断值。该分类器取得了不错的分类准确率；如果选择不同的截断值，其表现还能超过 DART。我们认为，这得益于它的简单性，而且数据量应足以构建潜在分布的合理核密度估计。

图18 展示了贝叶斯分类器的交叉验证兰德指数随截断值的变化情况。结果表明，使兰德指数达到最大值的最大截断值为 .288。（之所以选择最大的截断值，是因为在使兰德指数达到最大值的同时，这对应于最少的变量数。）图19 展示了在完整训练集（即使用完整数据集一半数据的训练集）上训练的贝叶斯分类器的兰德指数。该图显示了交叉验证截断值用于筛选完整训练集中的变量时的表现。图18 展示了所选变量数与截断值之间的关系。最后，图19 还显示了兰德指数对所选截断值的敏感程度。

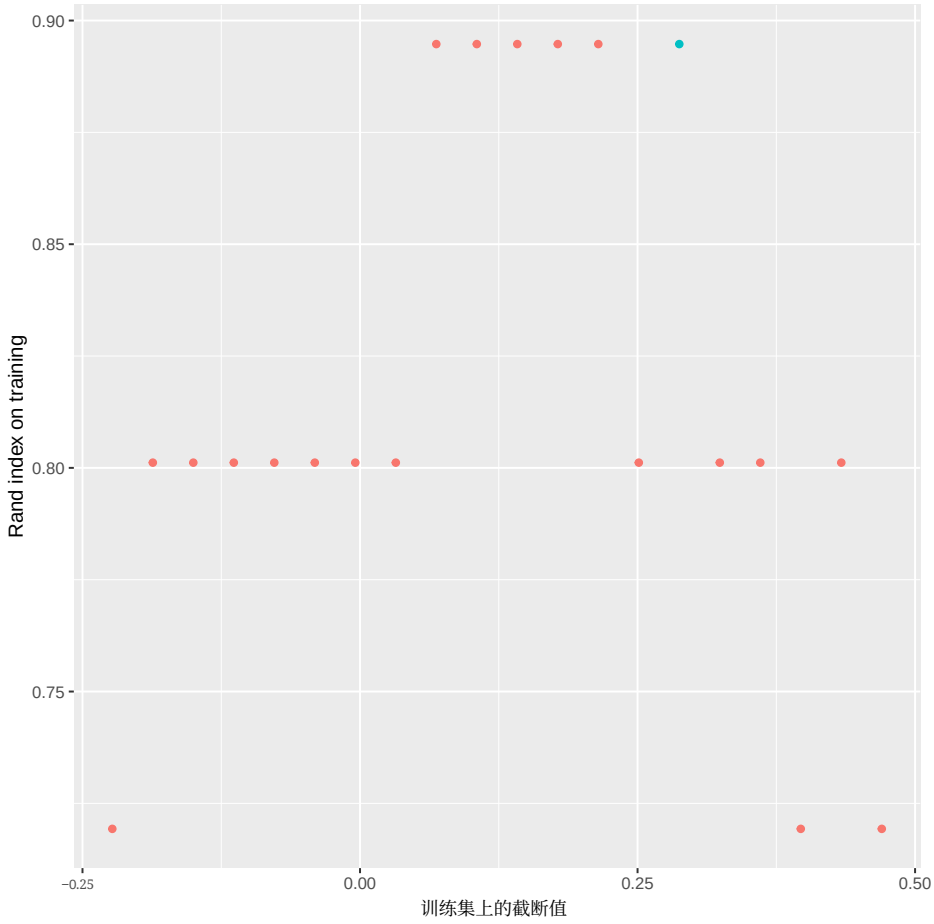


图18： 贝叶斯分类器在白血病数据训练集上的交叉验证性能。 方法对应的兰德指数图（5）与该方法选出的变量数之间的关系。兰德指数表示分类器在一个训练集上训练后应用于其他训练集时的性能。我们通过选择（蓝色）截断值来确定效果最佳的截断值：在使兰德指数达到最大值的一系列截断值中，它保留的变量最少。

我们认为，在其他一些情形下，贝叶斯分类器也有应用潜力：这些情形中，数据可能不足以训练集成方法，而且类别之间可能存在非位置类型的差异。我们注意到，与其他方法一样，使用贝叶斯分类器前最好先进行筛选，因为如果存在大量无用变量，所有这些方法的性能都会下降。令人鼓舞的是，在兰德指数大幅提升之前，0 是贝叶斯分类器对应的最大截断值；但稍显遗憾的是，交叉验证方法未能为该分类器选出能带来最佳性能的截断值

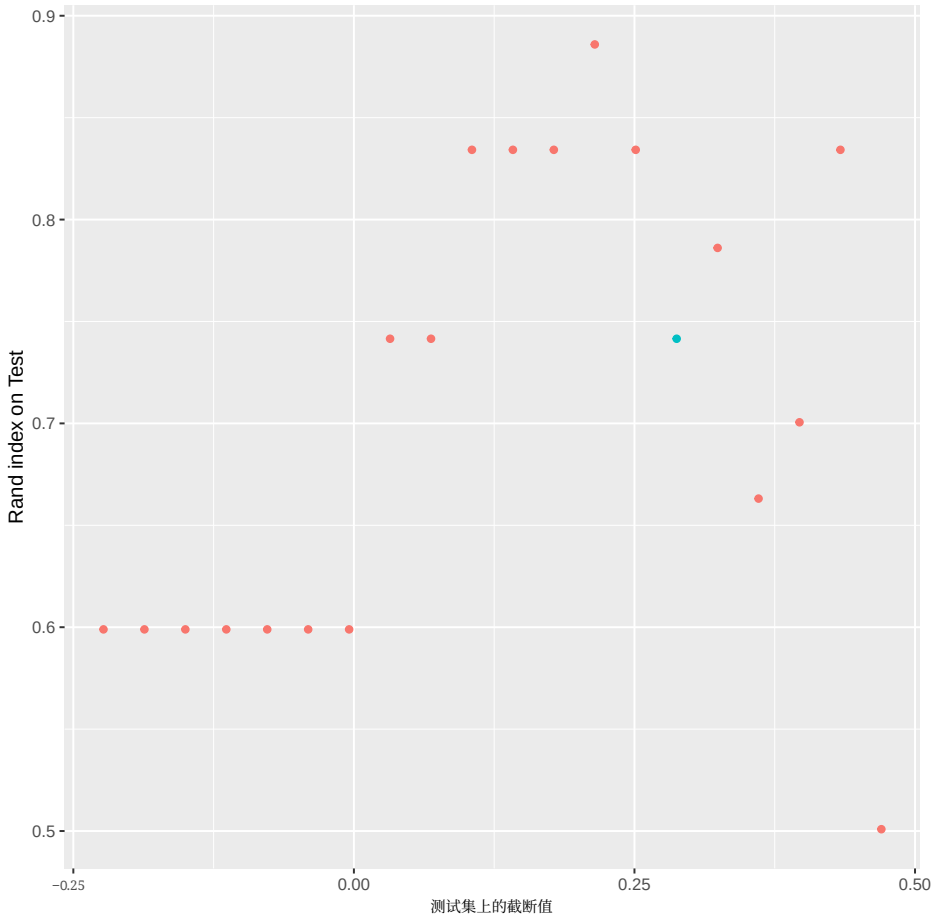


图19： 贝叶斯分类器在白血病数据测试集上的交叉验证性能。 兰德指数基于验证集。我们为交叉验证选择的截断值以蓝色表示?? ，该值是通过查看图 18确定的。

。问题在于，交叉验证方法选出的最优截断值对应于仅使用全部数据四分之一构建分类器的情形，而我们实际上需要知道使用一半数据时的最优截断值。未来研究可以关注最优截断值如何随训练集大小而变化。如果这种依赖关系足够简单，就有可能根据较小训练集大小下的交叉验证结果，估算给定训练集大小对应的最优截断值。

使用 ALB 筛选时，如果采用 k 具有 k 最大 ALB值的变量，效果往往更好。图21 展示了使用这种筛选方法后再应用 DART 得到的兰德指数。当变量数少于 500 时，ALB 筛选的效果优于类似的t-检验筛选方法。纳入的变量数足够多后， t-检验筛选的效果优于 ALB 筛选；但此时纳入的变量数已经足够多，以至于两种筛选方法得到的兰德指数都不理想。

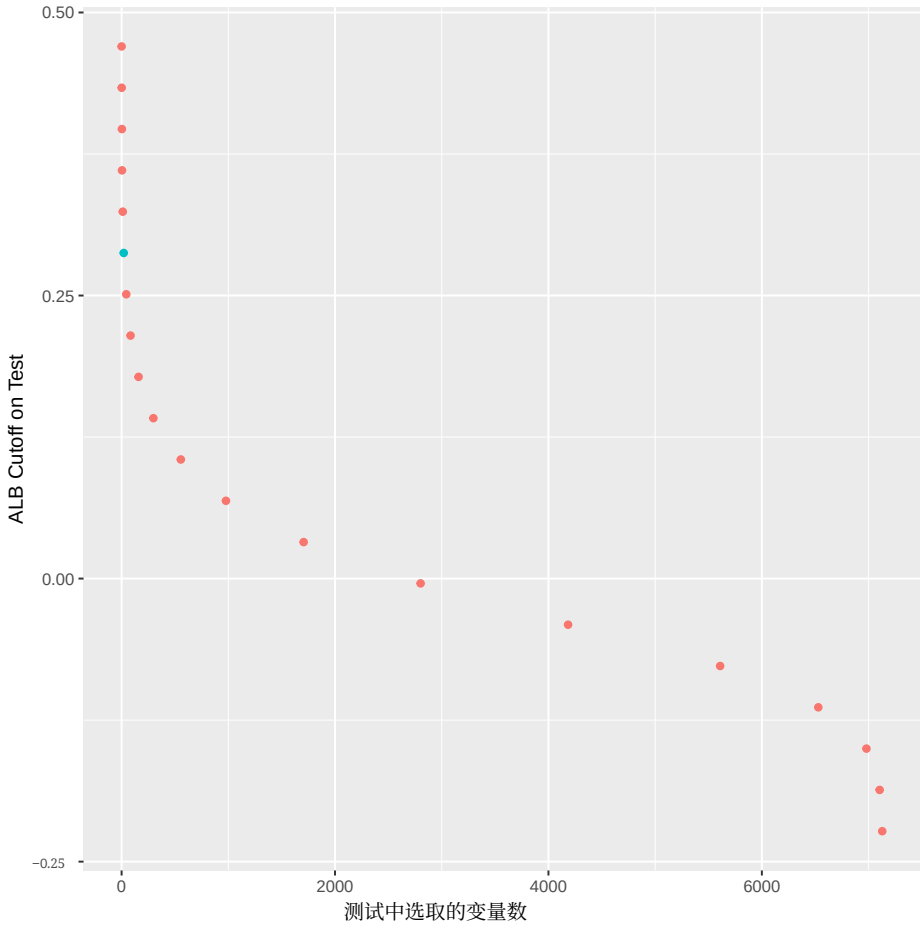


图20： 使用 *ALB*筛选方法在 *Leukemia*测试数据上，选取的变量数与所选截断值的关系。

9 结论与未来工作

我们提出了一种新的筛选方法，用于寻找位置类型以外的差异。要使这种方法比 t-检验筛选更有效，就需要将其与能够利用这些差异的分类方法结合起来。在模拟中，我们将 ALB 筛选与 BART、DART 和贝叶斯分类器结合，结果表明，当类别差异不属于位置类型时，其效果优于 t-检验筛选。将贝叶斯分类器应用于白血病数据集时，其表现优于 DART。即使数据主要包含位置差异，ALB 筛选的表现也不错，不过正如预期，它不如 t-检验筛选。

未来的工作包括努力提高计算 ALB 统计量及其置换分布的速度，尤其是针对大型数据集。SIS 已有一种筛选方法的迭代方法，未来研究可以探讨一种能够捕捉联合分布差异的 ALB 程序。本文中的模拟数据均采用了独立数据，而该方法对独立性的敏感程度也是值得探究的一个特性。ALB 筛选与随机投影或草图方法之间的相互作用，也是未来研究的一个有前景的方向；这些方法用于处理以下情况：

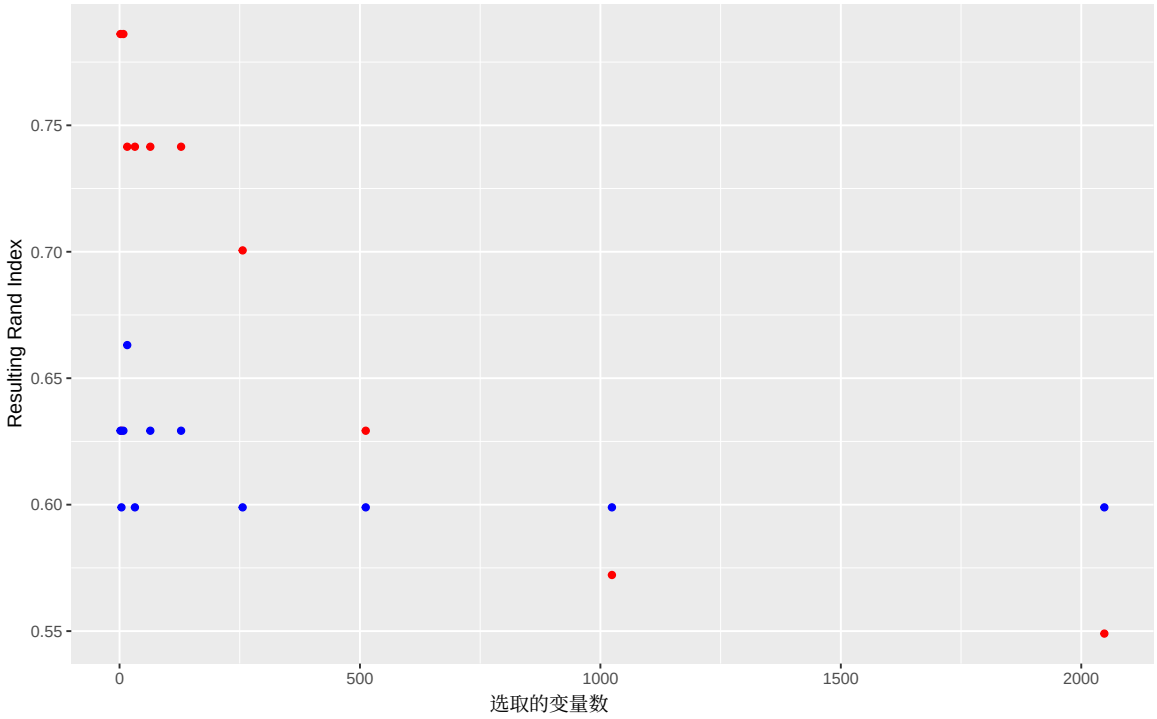


图21：白血病数据集的 *DART* 兰德指数与选取的变量数之间的关系。我们绘制了 DART 方法的兰德指数，其中相关变量是根据排名靠前的 ALB 或 t-统计量选取的。我们改变选取的变量数。红色点表示使用排名靠前的 ALB 的 DART 模型的兰德指数，蓝色点则表示使用排名靠前的 t-统计量的 DART 模型的兰德指数。

n 和/或 p 非常大。

参考文献

[Boser 等， 1992] Boser、B. E.、Guyon、I. M. 和 Vapnik、V. N.（1992年）。最优间隔分类器的训练算法。载于 计算学习理论第五届年度研讨会论文集，第 144–152 页。 1， 2

[Brieman 等， 1984] Brieman、L.、Friedman、J. H.、Olshen、R. A. 和 Stone、C. J.（1984年）。分类与回归树。 *Wadsworth Inc*，67。 1

[Cui 等， 2015] Cui, H., Li, R. 和 Zhong, W.（2015年）。超高维判别分析的无模型特征筛选。《美国统计学会会刊》，110(510)：630–641。1，2

[Fan 和 Lv， 2008] Fan, J. 和 Lv, J.（2008年）。超高维特征空间的确定性独立筛选。《英国皇家统计学会会刊：*B*辑（统计方法学）》，70(5)：849–911。2，4，11

[Fan 等人， 2010] Fan, J.、Song, R. 等人（2010年）。具有 NP 维度的广义线性模型中的确定性独立筛选。《统计年刊》，38(6)：3567–3604。2

[Friedman， 2002] Friedman, J. H.（2002年）。随机梯度提升。《计算统计与数据分析》，38(4):367–378。1

[Hill 等人， 2020] Hill, J.、Linero, A. 和 Murray, J.（2020年）。贝叶斯加性回归树：回顾与展望。《统计及其应用年度评论》，7。11

[Linero 和 Yang， 2018] Linero, A. R. 和 Yang, Y.（2018年）。能够适应平滑性和稀疏性的贝叶斯回归树集成。《英国皇家统计学会会刊：*B*辑（统计方法学）》，80(5):1087–1110。1，2，11

[Mai 和 Zou， 2012] Mai, Q. 和 Zou, H.（2012年）。用于高维二分类变量筛选的柯尔莫哥洛夫滤波器。《生物统计学》，100(1):229–234。1，2

[Merchant 和 Hart， 2022] Merchant, N. 和 Hart, J.（2022年）。一种基于核密度估计的贝叶斯启发式双样本检验。《熵》，24:1071。2，3，5，6

[Tang 等人， 2014] Tang, J.、Alelyani, S. 和 Liu, H.（2014年）。用于分类的特征选择：综述。《数据分类：算法与应用》，第 37 页。2

[Zhu 等人， 2011] Zhu, L.-P.、Li, L.、Li, R. 和 Zhu, L.-X.（2011年）。超高维数据的无模型特征筛选。《美国统计学会会刊》，106(496):1464–1475。4，11