 DOI: https://doi.org/10.26549/bdai.v4i3.13873

Clustering Knockoff Method for Controlling the Selection of High Dimensional Variables in FDR

Liexiao Ding

Big Data Analytics Trading Inc., America, 30067

Abstract

When studying high-dimensional data, the latitude of the data will be reduced to ensure that it is within the computable range before proceeding with the next calculation. Therefore, how to enhance the accuracy of variable selection is a problem that researchers need to solve. Therefore, researchers used Knockoff method to control the error detection rate, and compared with traditional BH method, its testing effect is better, achieving qualitative breakthroughs in the field of multiple tests. However, this method is limited to low-dimensional data and seriously affects the effectiveness of data control. Based on this, this paper proposes a clustering Knockoff method based on the above content, optimizes the operation process, and applies it to the selection of high-dimensional data variables to improve the rationality of the selection.

Keywords

multiple tests; Knockoff clustering method; FDR

聚类 Knockoff 方法控制 FDR 的高维变量选择

丁烈骁

Big Data Analytics Trading Inc.，美国 30067

摘　要

高维数据研究时，会降低数据纬度，保证其在可计算范围，才能进行下步计算，所以如何加强变量选择的准确性，是研究人员需要解决的问题。因此，研究人员利用Knockoff方法控制错误发现率，对比传统BH方法，其检验效果更佳，在多重检验领域中取得质的突破。但这种方法局限于低维数据，严重影响到数据控制效果。基于此，论文根据上述内容，提出聚类Knockoff方法，优化操作流程，将其应用到高维数据变量选择方面，提高选择的合理性。

关键词

多重检验；聚类Knockoff方法；FDR

1 引言

近年来，统计学覆盖到各行业，其最常用假设检验方法进行统计推断，即是提前收集各种样本数据，判断某个相关模型系数架设是否满足要求，根据假设数量将其分为多重检验和单个检验方法。其中单个检验存在时间较久，其应用方法较为完善，主要作用是控制第一类错误率，尽可能降低第二类错误率。而多重检验研究内容相对较少，通常被应用在医学领域、生物学领域等方面，如我们在研究某种药剂剂量时，既要分析计量的合理性，还要保证药效的安全性。但值得注意的是，随着生物基因领域研究进程不断深入，很多高维数据出现在我们视线内，给各种假设检验带来严峻挑战，我们要从大量基因中找到争取额数据，要综合考虑显著性判断和整体判断错误概率，不是单纯检验错误概率，即是进行多重检验 [1]。

【作者简介】丁烈骁（1987–），男，中国北京人，硕士，工程师，从事计算机网络架构研究。

2 聚类 Knockoff 方法框架

2.1 变量聚类

虽然高维统计建模不断发展，但仍然存在各种问题，如线性模型在高维数据领域中受到各种外在因素限制，当样本数量低于维度数量时，协方差矩阵为非满秩，导致无法求逆计算高维数据矩阵。针对该种问题，论文利用变量聚类方法，先聚类所有变量，再将不同类别的变量数量控制在样本数量内容，从而得到无数个低维样本的可逆协方差矩阵，通过将这些矩阵进行组合，形成完整的协方差矩阵。目前，变量聚类主要目的是增加组中变量之间的系数，降低组间相关系数。架设原始样本变量在 m 组，每组变量数量为（p1，p2，...，pm）。在应用过程中，我们通过分析各种生物数据，发现其分组特征较为明显，表示这种分组聚类具有较强合理性，拥有实践研究的作用。

2.2 分组构造 Knockoff 变量

对于每个分组，工作人员要根据原始变量数据，建设对应的分组中 Knockoff 变量，创建（n×（2pi））， 保证

设计矩阵满足行业标准 [2]。同时，通过分析上述内容，发现在分析每组变量时，Knochoff 系数矩阵和原始数据的系统矩阵基本相同，且不同数量的交叉相关系数和原始系数基本相同。例如：第 i 组中涉及变量 Xj 和 Xk：

$$\forall j,k \in gi, j \neq k \quad (1)$$

结合上述公式内容，评估出如果 j 为零假设，即 $\beta_j = 0$ 时：

$$X^T{}_j y \xleftrightarrow{d} \overline{X}^T{}_j y \quad (2)$$

公式（2）能证明相同分布性质基本成立，工作人员要将该种正确结论进行应用，得到 H0 下的交换性质，对于无法使用原假设的变量子集。

3 数值模拟

3.1 模拟数据

论文模拟数据集样本数量 500，特征纬度是高维样本数据，纬度数为 1000 维。在数据集模拟过程中，其真实出现在模型中的特征数量主要包括 50、100、200 等变量数量，有利于工作人员准确检测出变量数量给模型效果造成的影响。并且所有变量全部属于 随机生成变量，真实变量系数为 2.5，其他变量系数为 0[3]。同时，论文自动生成完整的设计矩阵，每个样本都要符合独立的分布数据，保证 $\sum$ 结构满足分组分块假设要求。其中 p 为 0.75，工作人员要沿着对角线拼接每组相关系数，不同变量分组之间的相关系数为 0，标准化处理 X，确保所有变量的均值等于 0，方差为 1。模型数据设置为：

$$y = x\beta + \in, \varepsilon \in N(0, I_n) \quad (3)$$

同时，在建模过程中，工作人员要利用不同 P 值，计算出 FDR 值，掌握其真正检验效果。

3.2 对比方法

目前，在数据模拟对比时，通常采用两种方法，计算在不同参数数据中的 FDR 数值和检验效果。

3.2.1 两阶段 Knockoff 方法

第一，要将原始数据 n=500 应用到 screening 变量筛选方程，根据行业标准选出前 80 个变量；第二，将上述变量传输到 Knockoff 方法计算框架，得到 $\overline{X} = X\left(1 - \sum^{-1} diag\{s\}\right) + \overline{UC}$。通过上述方法计算出 X 和 $\overline{X}$ 值，再将该数据应用到 LASSO 变量选择模型，计算出对应的统计量和阈值 T，找到 T 对应的筛选结果集 S，得到 FDR 和检验效果。

3.2.2 聚类 Knockoff 方法

先变量聚类 1000 个变量，通过 K- 均值聚类和层次聚类方法，科学设置不同类型的变量数据，保证其数量低于 500，类型数量选择时要根据实验计算出的最低类间距离进行。经过工作人员计算发现，这两种不同方法所产生的变量分别是 14 和 10。同时，工作人员要按照第一个对比方法，分别计算全部簇的 Knockoff 变量，集中这些变量数据，得到 $n \times 2p$ 的设计矩阵 $[X, \overline{X}]$，将其归纳到 LASSO 模型。在选择 FDR 控制阈值时，q 数值控制在 0.1-0.9 范围，找到不同 q 值对应的筛选结果集 S，科学评估检验效果和 FDR 值 [4]。

3.3 模拟结果

通过对比两种方法检验数据，发现当 FDR 值相同时，对比两阶段 Knockoff 方法，两种聚类 Knockoff 方法检验效果更佳，且阳性变量和应用效果成正比。同时，对于两种不同的聚类方法，层次聚类 Knockoff 方法效果高于 Kmeans 聚类 Knockoff 方法。因此，想要保证试验数据满足分块结果，工作人员尽可能采用聚类 Knockoff 方法，不仅能降低 FDR 数值，还能挑选出大量真实变量 [5]。

4 实证分析

4.1 基因微阵列数据

基因微阵列数据是生物学、物理学研究中的重要环节，生物学家通常将该种技术应用在特殊有机体基因方面，分析其全基因组的表达水平。微阵列属于特殊的玻璃载玻片，DNA 分子通过有规律的方式固定在指定位置，一个基因微阵列上涉及上千个位点，每个位点上有几百万个副本，这些副本和基因中的 DNA 分子相互对应。目前，基因微阵列数据被普及到基因表达分析领域，全面分析和基因相关的重要疾病原因，如糖尿病、癌症等，发现这些特殊基因给疾病带来的影响，促进药物研发工作能顺利进行。随着现代生物基因学快速发展，基因微阵列技术趋于成熟，但想要完美收集基因微阵列数据难度系数较高。主要原因其基因微阵列系数具有超高维性质，需要收集海量的基因位点，加上该技术使用成本较高，能收集的样本数量较少，易出现特征数量远超样本数量的问题，影响到其他分类算法、模式识别算法应用效果。针对该种问题，学者通常采用先降维后建模方法，其能有效减少计算复杂性，提高模型处理效率。同时，能将很多先进算法应用到该领域，如贝叶斯方法、SVN、决策树、神经网络等算法。但这种方法具有较强缺陷，在第一次降维中时常删除有用变量，所以如何降低降维中信息损失，是目前统计学研究的重要话题。

4.2 数据说明和处理

论文采用小圆蓝细胞瘤的基因表达谱作为数据集，其主要出现在儿童群体，是典型的恶性肿瘤。在该样本数据中有 2308 个基因数据，样本数量为 83 个，主要包括 EWS、BL、NB、RMS 等形态，对于不同形态所提出的治疗方法存在较强差异性，工作人员要利用基因微阵列数据判断出疾病类型。而该数据集是生物统计中最常见的小样本超高维数据集，如果直接将原样本应用到分类模型，会增加分类算法的复杂性，无法真正找到正确的基因序列。针对该种问题，工作人员通常利用三个步骤进行处理：第一步。变量选择原始数据；第二步，将降维后的数据应用到各种分类模型，为了准确评估肿瘤基因治疗序列，要尽可能压缩最终选择的基因集合，保证样本分类的准确性（如表 1 所示）。

表 1 SRBCT 数据样本组成

类别	原始数据集	训练集	测试集
EWS	29	23	6
RMS	25	20	5
NB	18	12	6
BL	11	8	3
合计	83	63	20

在研究上述数据集时，工作人员利用神经网络算法分类 96 个特征基因，将测试集分类准确率提升到 100%。同时，采用收缩质心算法选择 43 个特征因素，保证分类准确率为 100%，这是目前研究中使用特征数量最低的算法。

4.3 方法对比和分析结果

在由于论文使用的变量选择方法无法看到直观结果，主要原因其该数据集研究内容较多，我们很难找到基因数据中的有用特征。因此，我们要通过特征选择第二步分类算法进行评估，利用相同分类算法，根据相同分类数据达到同等分类准确率所需特征数量进行判断（如表 2 所示）。

表 2 多种方法对 SRBCT 数据的分类准确率

	特征选择方法	特征个数	分类方法	分类准确率
Khan	灵敏度	96	神经网络	100%
Yeo	信噪比	60	支持向量机	100%
Tibshirani	收缩后的质心距离	43	K 近邻	100%
Barberh Candes	两阶段 Knockoff	15	支持向量机	65%
本研究	聚类 Knockoff	15	支持向量机	100%

当分类准确率相同时，Tibshirani 选择 43 个特征基因，将其应用到分类模型中，得到 100% 的分类准确率，是目前应用各种数量最少的方法；但在论文分类检验中，要保证 FDR 数据低于 0.3，变量聚类 2308 个变量，得到 182 类最佳聚类数量，计算每个类别变量产生的 Knockoff 变量，再将这些变量合并。然后将其全部应用到 LASS 变量选择模型，选择出 15 个变量，得到 100% 分类准确率。论文也应用两阶段算法，但如果将 FDR 控制在 0.3 内，只能选择 5 个变量，分类准确率为 65%。因此，站在分类准确率角度来看，本文方法高于两阶段方法。

5 结语

综上所述，多重检验是解决上述问题的主要方法，是通过合理控制错误发现率、整体错误率等数据，控制第一类错误率，保障两者概率低于行业标准值，让工作人员合理选择变量。基于此，研究人员利用 Knockoff 方法对错误发现率进行控制，其测试效果更好，与传统 BH 方法相比，在多重测试领域实现了质的突破。但这种方法被限制在数据控制效果受到严重影响的低维数据上。根据以上内容，论文提出了在高维数据变量选择中应用聚类 Knockoff 方法，对操作流程进行了优化，提高了选择的合理性。

参考文献

[1] 王小燕,张中艳.基于Knockoff的分位数回归变量选择方法及其投资组合决策应用[J].统计研究,2023,40(4):124-137.

[2] 李泓.基于FDR控制的函数型数据的变量选择研究[D].兰州:兰州大学,2022.

[3] 王小燕,周颖,唐婷婷,等.基于Knockoff-Logistic的多因子量化选股研究[J].统计与信息论坛,2023,38(4):19-32.

[4] 赵学斌.基于Knockoff框架的变量选择模型构建与分析[D].武汉:华中农业大学,2022.

[5] 夏亚峰,何佳.高维数据下广义线性模型自适应桥惩罚估计的变量选择[J].甘肃科学学,2022,34(1):9.