存在混杂时高维数据的随机森林分析

尤东方,魏永越,张汝阳,陈 峰,赵 杨*

南京医科大学公共卫生学院生物统计学系,生物医学大数据重点实验室,江苏 南京 211166

[摘 要] **目的**:探讨存在混杂因素时高维数据中随机森林(random forest,RF)的分析方法。**方法**:通过模拟实验和实例数据分析对单纯随机森林分析、增加节点候选变量为最大值以及基于广义线性模型的残差校正混杂因素的结果进行比较,以重要变量的重要性评分排序情况进行评价。**结果**:模拟实验表明,增加节点候选变量的方法对混杂因素的校正效果不明显,而基于广义线性模型残差的方法能有效校正混杂效应;实际数据分析结果显示单纯随机森林分析rs3754686和rs2322660分别排在第一和第二位。增加节点候选变量后rs3754686排序变化较小,而基于残差的方法校正人群分层后这两个单核苷酸多态位点(SNPs)的排序大幅度降低,从而打破乳糖酶(LCT)基因与身高之间的虚假关联。**结论**:随机森林分析需要考虑混杂因素问题,基于广义线性模型的残差能有效校正混杂因素,适用于高维数据的变量筛选。

[关键词] 随机森林;混杂因素;残差;人群分层

[中图分类号] O212 [文献标志码] A [文章编号] 1007-4368(2018)07-978-05

doi:10.7655/NYDXBNS20180720

A random forest analysis of high-dimensional data with the confounding effects

You Dongfang, Wei Yangyue, Zhang Ruyang, Chen Feng, Zhao Yang*

Department of Biostatistics, School of Public Health, Key Laboratory of Biomedical Bigdata, NMU, Nanjing 211166, China

[**Abstract**] **Objective**: This project explored a random forest (RF) analysis of high-dimensional data with the confounding effects. **Methods**: We used computer simulations and real data validation to evaluate the performance of 2 methods which can potentially account for the confounding effects in RF analysis: RF analysis with maximum candidate variables at each split (RFMCV) and RF with glm-based correction. The distribution of ranks of the causal variable was used to evaluate these approaches. **Results**: Simulation experiments suggested that RF with glm-based correction was more effective than the RFMCV to correct the confounding effects. The real data validation showed that rs3754686 and rs2322660 were ranked first and second, respectively. Analysis results of GWAS data confirmed that RF with glm-based correction can effectively remove the spurious association between the LCT gene and height. **Conclusion**: The confounding effects should be correctly adjusted in RF analysis. RF with glm-based correction was applicable to adjust the confounding effects and variable selection in high-dimensional data.

[**Key words**] random forest; confounding effect; residual; population stratification

[Acta Univ Med Nanjing, 2018, 38(07): 978-982]

随着高通量技术的飞速发展,高维组学数据的分析成为了热点和难题[1],传统的统计学方法如t检验、卡方检验和非条件logistic回归等的使用受到了限制。近年来,大量研究表明随机森林(random forest,RF)作为一个有效的机器学习方法,能较好地处理高维组学数据[2-5]。RF既可以用于分类也可以用于回归,不但能对结局进行预测,且能够提供每个变量的重要性大小,从而可用于高维数据的变量筛选,在组学数据分析中得到了日益广泛的应用。

然而大部分的生物医学数据来自于观察性研

[基金项目] 国家重点科研项目(2016YFE0204900);国家自然科学基金(81373102, 81530088, 81473070, 81402764, 81402763);江苏省青蓝工程学科带头人;江苏省预防医学优势学科;江苏高校品牌专业建设工程资助项目(PPZY2015A067);江苏省自然科学基金重点项目(14JA31002)

*通信作者(Corresponding author),E-mail:zhaoyang@njmu.edu.cn

究，易受到混杂因素的干扰。在传统的多因素分析中，往往通过多因素回归分析，将混杂因素作为协变量以进行调整。因而，部分研究者在利用随机森林模型进行多因素分析时，也视混杂因素与所关心的研究因素地位相同，将两者等同对待，认为这样能调整混杂因素的干扰。

本文将首先说明这种做法存在缺陷，然后再利用模拟实验和真实案例，比较两种可能的混杂因素调整方法。

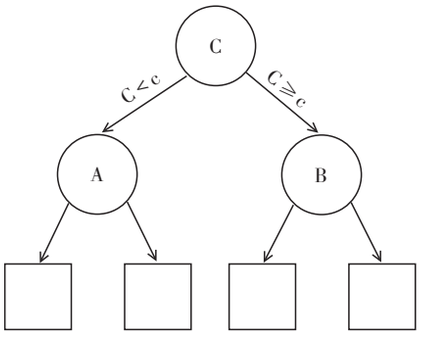
1 原理与方法

1.1 随机森林

RF是利用bootstrap重采样的方法从原始数据中抽取多个样本，称之为bootstrap样本，然后对这些样本分别构建决策树，组合多棵决策树构成最后的随机森林。对于每棵决策树的生长，每个节点都将随机从M个变量中随机选取m($m \leqslant M$)个进行分割，即节点候选变量，而不是通过所有变量来决定，从而引入随机性，提高预测精度。而节点候选变量数的大小影响着RF的强度和相关性，从而影响着RF的预测精度。RF在进行分析的时候，同时给出变量重要性评分(variable importance measure，VIM)，用来衡量不同变量对模型的贡献量，变量重要性评分越大说明该变量对模型的预测越重要。RF的变量重要性评分有两种，分别是通过基尼(Gini)指数和变量值的置换方法计算其重要性[6-7]。有文献报道置换法比基尼指数法具有更小的偏倚性，尤其在处理高维数据时[4,8]。本文在分类(classification)模型中选择MDA(mean decrease in accuracy)为VIM统计量，回归(regression)模型中则选取IMSE(increased mean square error)。

1.2 传统方法无法在RF分析中调整混杂因素

医学研究中，混杂因素往往会掩盖或夸大研究因素和结局之间的真实关联性，如何有效消除混杂因素的影响至关重要。传统做法有协方差分析、分层分析以及多因素回归分析。此类做法均是相当于把混杂因素固定在某个水平上来评价研究因素对结果的影响。而在RF中，这就必须要求混杂因素位于高位节点。如图1所示，当混杂因素C为根节点时，随机森林按某个值将其分为两组，分别在不同的组下继续评价研究因素的效应，这样即为将混杂因素固定在某个水平上从而达控制混杂效应的作用。然而在各变量竞争成为节点时，并不能保证混杂因素在每棵树中都能较早被选中，更不能保证混杂因素一定在重要变量的父节点上。此时，如果按传统做法直接将混杂因素与各研究变量一同构建随机森林，将不能起到调整混杂因素的作用。



A、B：混杂因素的子节点；C：混杂因素；c：混杂因素节点分割值。

图1 随机森林图示

Figure 1 Overview of random forest

1.3 通过增加节点的候选变量数调整混杂

为了提高预测精度，RF在节点变量的选取上引入随机性，即每个节点的候选变量是从所有M个变量中随机选取一部分。针对节点候选变量数，随机森林对于不同模型有不同的默认值(分类模型为$\sqrt{M}$，回归模型为$M/3$)，在实际研究中也可自行设置大小。因此，部分研究者认为，通过增加节点的候选变量数，从而增加混杂因素被较早选中的概率来调整混杂因素。为了尽可能让混杂因素被选中，本研究将节点候选变量数设置为最大值M，即将所有变量都作为节点的候选变量，继而竞争成为节点。

1.4 基于广义线性模型的残差调整混杂因素

本文作者在利用RF模型分析全基因组关联性数据时，针对人群分层带来的混杂效应，提出了一种基于主成分分析和广义线性模型的方法[9]。其基本原理是：首先利用主成分分析获得主成分，用以量化人群分层效应大小，再利用各变量与获取的主成分构建广义线性模型，从而扣除人群分层带来的混杂效应。

具体做法为：首先将应变量与获取的主成分构建广义线性模型，得到模型的残差，也就是真实值与模型的拟合值之差，此时得到的残差即为调整后的应变量。同样的，针对各个自变量分别与混杂因素构建广义线性模型，相应的残差为调整后的各个自变量。

2 模拟实验

模拟实验将构建两种情况，分别设置不同的混杂因素，并加入众多噪声变量，考察单纯RF法、增加节点候选变量数来调整混杂因素的随机森林分析法的应用效果，同时与基于广义线性模型的残差法进行

比较。本研究随机产生100万例数据(包括病例和对照),形成模拟数据,然后从中随机选取疾病组(Y=1)和对照组(Y=0)的各1 000例构成样本。模拟重复1 000遍,观察单纯RF分析(节点候选变量数为默认值)、节点候选变量数为最大值以及基于广义线性模型的残差调整混杂因素时重要变量的排序情况。

模拟实验一设置1个重要变量、1个混杂因素以及99个噪声变量,均为二分类变量。其中重要变量和混杂因素通过R软件中的bindata包产生,相关系数为-0.4,随机产生的99个噪声变量服从伯努利分布。应变量Y通过logistic回归模型产生,模型为:$\log it(\pi_y)=-4+\ln(1.5)X_{causal}+\ln(2)C$,其中$X_{causal}$为重要变量,$C$为混杂因素。

模拟实验二同样设置均为二分类的1个重要变量和99个噪声变量,而混杂因素将增加至2个。其中重要变量和2个混杂因素(分别用C_1和C_2表示)的相关系数矩阵如下:

$$R=\begin{bmatrix} r_{xx} & r_{xc_1} & r_{xc_2} \\ r_{xc_1} & r_{c_1c_1} & r_{c_1c_2} \\ r_{xc_2} & r_{c_1c_2} & r_{c_2c_2} \end{bmatrix}=\begin{bmatrix} 1.00 & -0.60 & -0.50 \\ -0.60 & 1.00 & 0.30 \\ -0.50 & 0.30 & 1.00 \end{bmatrix}$$

产生应变量Y的Logistic回归模型如下:

$$\log it(\pi_y)=-4+\ln(2)X_{causal}+\ln(1.5)C_1+\ln(2)C_2$$

2.1 模拟结果

表1给出了单纯RF分析、节点候选变量数为最大值以及基于广义线性模型的残差调整混杂因素在模拟实验中重要变量的排序情况。模拟实验一显示存在1个混杂因素的情况下,节点候选变量数从默认值改变到最大时,重要变量按VIM排序排在第一的比例由0.5%提升至1.9%,而采用基于广义线性模型残差校正混杂的方法,重要变量分别有38.5%和65.2%的比例排在第1和前5位。模拟实验二结果显示当存在2个混杂因素时,节点候选变量数为最大值时重要变量排第1的比例从2.1%仅提升至4.0%,基于广义线性模型的残差调整混杂因素的方法能大幅度的提升至83.0%,且有97.0%的比例排在前5位。

图2和图3分别给出了只有1个混杂因素和2个混杂因素时重要变量在各个方法下VIM排在第1的分布情况。结果均显示通过增加节点候选变量数的方法对混杂因素的校正效力较小,而基于广义线性模型的残差法在随机森林分析中能有效地校正混杂因素,使得重要变量能被识别出来。

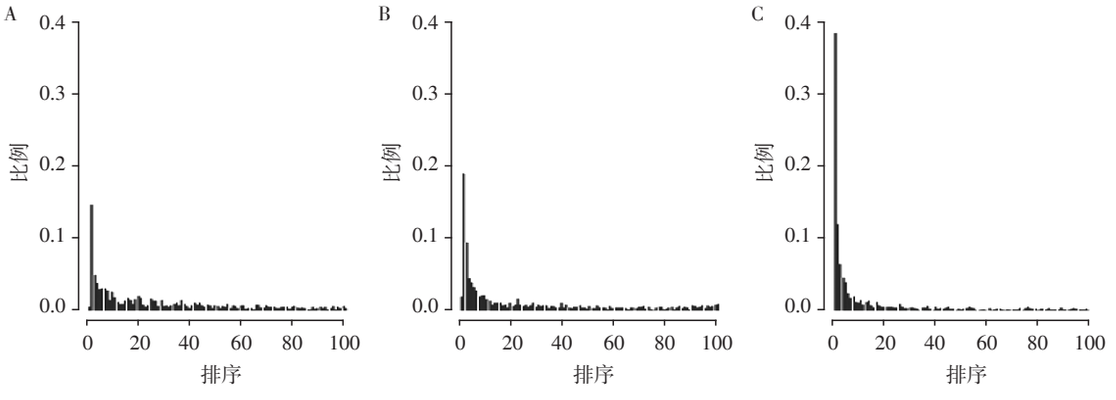
2.2 实例验证

多项研究报道由于人群分层(population stratification)所导致的LCT基因中的单核苷酸多态性(single nucleotide polymorphisms,SNPs)与身高之间的虚假关联[10-12]。本研究使用哈佛肺癌易感性研究(Harvard Lung Cancer Susceptibility Study)的全基因

表1 模拟实验中重要变量排序在第1、前5、前10位的比例

Table 1 Proportion of causal variables with ranks of 1, <=5, and <=10 in scenario 1 and 2

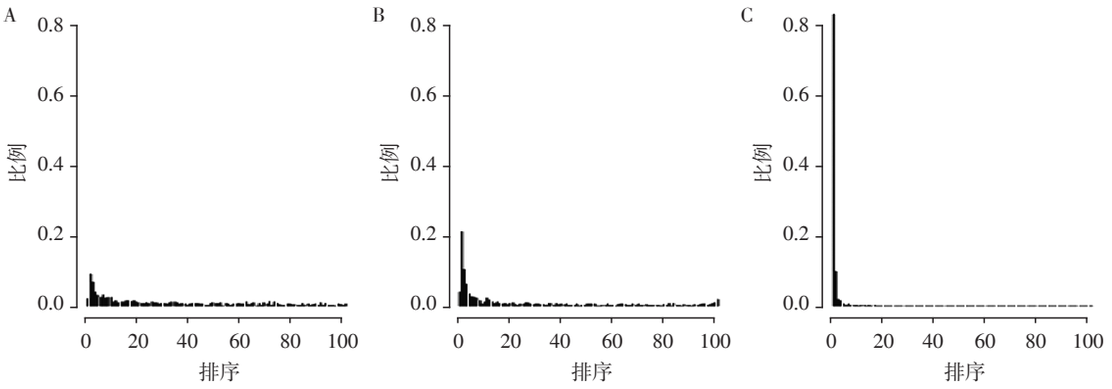
模拟实验	单纯随机森林法			节点候选变量数最大化			基于广义线性模型的残差法		
	排序=1	排序≤5	排序≤10	排序=1	排序≤5	排序≤10	排序=1	排序≤5	排序≤10
模拟实验一	0.005	0.267	0.393	0.019	0.386	0.506	0.385	0.652	0.733
模拟实验二	0.021	0.249	0.379	0.040	0.453	0.547	0.830	0.970	0.983



A:单纯随机森林;B:节点候选变量数最大化;C:基于广义线性模型的残差法。

图2 模拟实验一中重要变量排序分布图

Figure 2 Distribution of causal variables ranks from scenario 1



A：单纯随机森林；B：节点候选变量数最大化；C：基于广义线性模型的残差法。

图3　模拟实验二中重要变量排序分布图

Figure 3　Distribution of causal variables ranks from scenario 2

组关联性研究(genome - wide association study，GWAS)数据。DNA取自于全血且基因型分型使用Illumina 610 k Quad chip。该研究的具体描述见Zhao等[9]的文章。本研究选取了与LCT基因具有高连锁不平衡(linkage disequilibrium，LD)的SNP，分别为rs3754686和rs2322660。最终数据包含859个样本以及1 000个SNPs。本研究将身高按175 cm为界，划分为二分类变量用于随机森林模型的构建，分别对数据进行单纯随机森林分析(即节点候选变量数为默认值)、节点候选变量数为最大值以及基于广义线性模型的残差来调整人群分层带来的混杂效应，观察rs3754686和rs2322660的变量重要性评分排序情况。其结果显示单纯随机森林分析，即未对混杂因素作调整时，rs3754686和rs2322660的重要性评分在1 000个SNPs中分别排在第1和第2位。当节点候选变量数为最大值时，rs3754686和rs2322660的变量重要性评分排序在第28和第111位。当使用基于广义线性模型的残差调整人群分层时，此时rs3754686和rs2322660则分别排在第551和443位。上述结果显示未做任何调整的单纯随机森林分析，LCT基因的两个SNPs由于混杂因素人群分层的存在显示出了与身高的虚假关联；当节点候选变量数为最大值时，对人群分层有一定的调整作用，但是rs3754686的排序仍然较靠前，易被错误地筛选出来；而基于广义线性模型的残差调整人群分层后，rs3754686和rs2322660两个变量的重要性评分排序均非常靠后，从而打破了这两个SNPs与身高之间的虚假关联。

3　讨　论

随机森林作为一种机器学习方法，由若干决策树构成的组合模型，也是目前最精确的学习算法之一，能有效处理非线性、交互作用等问题，同时还能够防止过拟合，也可用于高维组学数据的预测和变量筛选。Hsieh等[13]比较了几种机器学习方法，发现相较于人工神经网络和支持向量机，随机森林的预测更准确。Pang等[14]的研究表明RF可以处理变量为多分类的情况，并进行回归分析，相比于判别分析等分类方法，具有更大的优势。

但当研究数据中存在混杂时，简单地将混杂因素当协变量一同放入RF中的分析方法并不能有效地校正混杂因素。本研究利用模拟实验，比较了当存在1个、2个混杂因素时，不进行校正、增加节点候选变量数和基于广义线性模型的残差调整法的分析结果。本研究的模拟实验可见，当节点候选变量数为默认值时，并不能有效筛选出重要变量，从而也验证了简单将混杂因素作为协变量的处理方法在RF分析中并不适用，这是因为并不能保证在RF中混杂因素总出现在重要变量的父节点上。如果增加节点候选变量将候选变量数设为最大值，即在每个节点处候选变量为所有变量，模拟实验结果显示该种方法校正混杂因素的效果并不明显，这是因为当两个变量竞争成为节点变量的时候，混杂因素效应较强的时候，重要变量的重要性可能会被削弱，从而不能有效地筛选出重要变量。而基于广义线性模型残差法校正混杂因素，则是通过各变量与混杂因素构建广义线性模型，用真实值减去拟合值得到残差，从而达到将混杂效应从各变量中分离出去的效果。模拟实验也显示不论是只存在1个混杂还是2个混杂因素，基于广义线性模型的残差法均能很好地校正混杂因素，继而再用RF分析能准确地筛选出重要变量。

本研究选用经典的由混杂因素人群分层的存在导致LCT基因与身高的虚假关联作为实例验证。直接将混杂因素作为协变量放入RF中rs3754686和rs2322660排在前2位，将节点候选变量数增加为最大值时，rs3754686的排序仍然较靠前，显示出该方法并没有很好地校正人群分层的效应。采用基于广义线性模型残差的方法，rs3754686和rs2322660的排序均非常靠后，从而打破了LCT基因与身高的虚假关联。从模拟实验以及实例验证的结果可见，增加节点候选变量数的方法不能有效校正混杂因素，基于广义线性模型残差的方法在随机森林中能有效调整混杂因素，从而正确地筛选出重要变量。

作为一种数据挖掘方法的RF已经在很多领域中得到较好应用，使用时应注意根据不同的数据设定相应的参数，如决策树的数量等。另外，本研究模拟实验只探讨混杂因素只有1个或2个的情况，更复杂的混杂因素情况还有待进一步研究。

[参考文献]

[1] 宋欠欠，李轶群，侯 艳，等. 随机森林的变量捕获方法在高维数据变量筛选中的应用[J]. 中国卫生统计，2015，32(1)：49-53

[2] Goldstein BA，Hubbard AE，Cutler A，et al. An application of pandom forests to a genome-wide association dataset：Methodological considerations & new findings [J]. BMC Genet，2010，11(1)：1-13

[3] Kim Y，Wojciechowski R，Sung H，et al. Evaluation of random forests performance for genome - wide association studies in the presence of interaction effects [J]. BMC Proc，2009，3(S7)：S64

[4] Nicodemus KK，Malley JD，Strobl C，et al. The behaviour of random forest permutation - based variable importance measures under predictor correlation [J]. BMC Bioinformatics，2010，11(1)：110

[5] Yan VS，Cai Z，Desai K，et al. Classification of rheumatoid arthritis status with candidate gene and genome-wide single - nucleotide polymorphisms using random forests [J]. BMC Proc，2007，1(S1)：S62

[6] Breiman L. Random forest[J]. Mach Learn，2001，45(1)：5-32

[7] 武晓岩，李 康. 随机森林方法在基因表达数据分析中的应用及研究进展[J]. 中国卫生统计，2009，26(4)：437-440

[8] Nicodemus KK，Malley JD. Predictor correlation impacts machine learning algorithms：implications for genomic studies[J]. Bioinformatics，2009，25(15)：1884-1890

[9] Zhao Y，Chen F，Zhai R，et al. Correction for population stratification in random forest analysis[J]. Int J Epidemiol，2012，41(6)：1798-1806

[10] Campbell CD，Ogburn EL，Lunetta KL，et al. Demonstrating stratification in a European American population [J]. Nat Genet，2005，37(8)：868-872

[11] Qin H，Morris N，Kang S J，et al. Interrogating local population structure for fine mapping in genome-wide association studies [J]. Bioinformatics，2010，26(23)：2961-2968

[12] Li M，Reilly MP，Rader DJ，et al. Correcting population stratification in genetic association studies using a phylogenetic approach [J]. Bioinformatics，2010，26(6)：798-806

[13] Hsieh CH，Lu RH，Lee NH，et al. Novel solutions for an old disease：diagnosis of acute appendicitis with random forest，support vector machines，and artificial neural networks[J]. Surgery，2011，149(1)：87-93

[14] Pang H，Lin A，Holford M，et al. Pathway analysis using random forests classification and regression [J]. Bioinformatics，2006，22(16)：2028-2036

[收稿日期] 2017-11-21

(上接第944页)

学，2004，9(9)：1069-1072

[24] 伍腾飞，彭秀军. EDTA的螯合作用在角膜研究中的应用进展[J]. 国际眼科杂志，2012，12(10)：1890-1893

[25] Najjar DM，Cohen EJ，Rapuano CJ，et al. EDTA chelation for calcific band keratopathy：results and long-term follow-up[J]. Am J Ophthalmol，2004，137(6)：1056-1064

[26] Hachem R，Bahna P，Hanna H，et al. EDTA as an adjunct antifungal agent for invasive pulmonary aspergillosis in a rodent model[J]. Antimicrob Agents Chemother，2006，50(5)：1823-1827

[27] Movahedian A，Zolfaghari B，Mirshekari M. Antioxidant effects of hydroalcoholic and polyphenolic extracts of *Peucedanum pastinacifolium* Boiss. & Hausskn [J]. Res Pharm Sci，2016，11(5)：405-411

[收稿日期] 2017-12-17