基于 AUC 统计量的随机森林变量重要性评分的研究*

哈尔滨医科大学卫生统计教研室(150081) 张晓凤 侯 艳 李 康△

随机森林(random forest,RF)[1]是高维组学数据常用的分析方法,在进行判别分析时,同时能够给出变量重要性评分(variable importance measure,VIM)。RF 的变量重要性评分通常有两种,一种方法是通过变量值的置换计算其重要性,第二种方法是通过基尼(Gini)指数计算其重要性,由于置换法比 Gini 指数法具有更好的非偏倚性能,因此多采用置换法进行变量筛选[2-5]。然而,当数据类别(标签)比例不均衡时,即收集到的数据在两类中的数目不相同,尤其比例相差较大时,基于错误率(error rate,ER)的置换法不能准确反映变量的重要性。为此,Janitza 等(2013)提出基于 *AUC* 统计量的评价方法,能够克服类别间比例不平衡的影响[6]。本文在简要介绍该方法的基础上,通过模拟实验和实例数据探索其适用性,并与传统的置换法进行比较。

原理与方法

1. RF 的基本思想

RF 采用组合方法(ensemble method)的思想,即对样本数据进行多次随机抽样产生 N(通常为 *Ntree*)个训练样本构造 N 棵分类树(称基分类器),在每次基分类器构建过程中,将训练样本以外的数据作为测试数据,称为袋外数据(out of bag data sets,OOB),并通过错误率来评价基分类器性能,最后根据投票(vote)准则将基分类器组合为一个 RF 分类器。RF 在构建分类器的过程中,通过对变量重要性排序进行变量重要性评分。

2. 基于错误率的置换方法

基于错误率置换方法的变量重要性评分(VIM_ER),其基本原理是用同时随机置换各变量值,通过计算置换前后的 OOB 错误率间的差异衡量该变量的重要性。具体地,欲获得变量 X_i 的重要性评分,首先基于训练样本构建随机森林,并估计所有 OOB 样本的错误率,然后对所有 OOB 样本中的变量 X_i 值进行打乱获得新的袋外数据(OOB′),估算 OOB′样本的 ER,最后计算两次袋外数据的 ER 变化值。最后将所有 OOB 样本 ER 变化均值作为 X_i 的 VIM,X_i 的 VIM 定义如下:

$$VIM_i^{ER} = \frac{1}{Ntree}\sum_{t=1}^{Ntree}(ER_{it} - ER'_{it}) \qquad (1)$$

其中,*Ntree* 为 RF 中树的个数,ER_{it} 为变量 X_i 置换之前第 t 棵树对应的错误率,ER'_{it} 为变量 X_i 置换之后第 t 棵树对应的错误率。

由 VIM 计算公式我们知道,如果变量 X_i 与标签(类别)无关联,随机置换该变量后对应的袋外数据错误率不会发生变化,理论上 $VIM_i^{ER}=0$;相反地,如果 $VIM_i^{ER}>0$,则说明变量 X_i 与分类是有关联的。

3. 基于 *AUC* 统计量的置换方法

基于 *AUC* 统计量置换法同样能够得到变量的重要性评分(VIM_AUC),与 OOB 错误率得到的 VIM_ER 原理相似,两者区别在于后者基于错误率变化衡量变量重要性,前者则是基于 *AUC*(ROC 曲线下面积)值的变化评价变量重要性。这里,变量 X_i 重要性评分定义如下:

$$VIM_i^{AUC} = \frac{1}{Ntree}\sum_{t=1}^{Ntree}(AUC_{it} - AUC'_{it}) \qquad (2)$$

其中,AUC_{it} 为变量 X_i 置换之前第 t 棵树对应的 *AUC* 值,AUC'_{it} 为变量 X_i 置换之后第 t 棵树对应的 *AUC* 值。

使用 OOB 错误率的变化作为评价变量重要性的指标时,考虑的是整体错误率变化情况,但最大的问题是当多数类样本较大时,OOB 错误率未充分考虑少数类的错误率,相当于赋予了多数类更高的权重。基于 *AUC* 统计量的置换方法同时考虑灵敏度和特异度,相当于对两类各自的准确率赋予了相同的权重,直观上,对于类别间不平衡数据而言,基于 *AUC* 统计量得到的变量重要性评分更趋于合理。

模拟实验

1. 实验目的

(1)探索处理不平衡数据时基于 ER 估计 VIM 的偏倚性,验证基于 *AUC* 统计量获得 VIM 的合理性。

(2)比较 VIM_ER 和 VIM_AUC 对变量排序的差别,以及对差异变量和噪音变量的区分能力。

2. 实验设置

(1)模拟数据共设置 65 个自变量 $X=(X_1,\cdots,X_{65})$ 和一个应变量 $Y\in\{0,1\}$,其中按自变量与应变量

* 基金项目:国家自然科学基金资助(81473072)
△通信作者:李康,E-mail:likang@ems.hrbmu.edu.cn

之间的关联程度设置强、中、弱、无四个等级，共 15 个变量，称为差异变量；另外设置 50 个无关联变量，称为噪音变量，具体分布情况见表 1。现设置，分组 1 为样本较少一组，分组 2 为样本较多一组；两组类别样本量不平衡的比例(n_1:n_2)为 1:1，1:3，1:5，1:10，1:15，1:20；第一组的样本含量分别为 10 和 30，实验重复 100 次。

(2)随机森林构建参数设置，分类树 *Ntree* = 1000，*mtry* = 5，基分类器构建时抽取的训练数据为无放回抽样。

表 1　自变量的分布参数设置

特征(变量)	分组 1 分布	分组 2 分布	差异大小
$X_1,\cdots,X_5$	N(1,1)	N(0,1)	强
$X_6,\cdots,X_{10}$	N(0.75,1)	N(0,1)	中
$X_{11},\cdots,X_{15}$	N(0.5,1)	N(0,1)	弱
$X_{16},\cdots,X_{65}$	N(0,1)	N(0,1)	无

3. 模拟实验结果

(1)图 1 和图 2 分别给出两组样本量平衡和不平衡情况下，VIM_ER 和 VIM_AUC 两种方法的结果。图 1 结果显示，在两组例数相同时，VIM_ER 和 VIM_AUC 两种方法均能真实反映变量重要性；图 2 结果显示，在两组例数不相同、并且相差较大时(n_1:　n_2 = 1:20)，VIM_ER 方法几乎看不到差异变量的作用，而 VIM_AUC 方法能更好地区分出差异变量，比 VIM_ER 方法更合理。

(2)图 3 给出了两组样本量不相同情况下，VIM_ER 和 VIM_AUC 两种方法区分差异变量的能力。结果显示，随着两组不平衡比例增加，VIM_ER 法对差异变量区分的 *AUC* 值呈下降趋势，表明两组样本比例不平衡时，VIM_ER 方法获得的变量 VIM 得分不能很好地识别差异变量；而 VIM_AUC 法得到的 *AUC* 值随着总样本量的增加而增加，最后趋于稳定，表明 VIM_AUC 不受两组样本例数不平衡的影响。

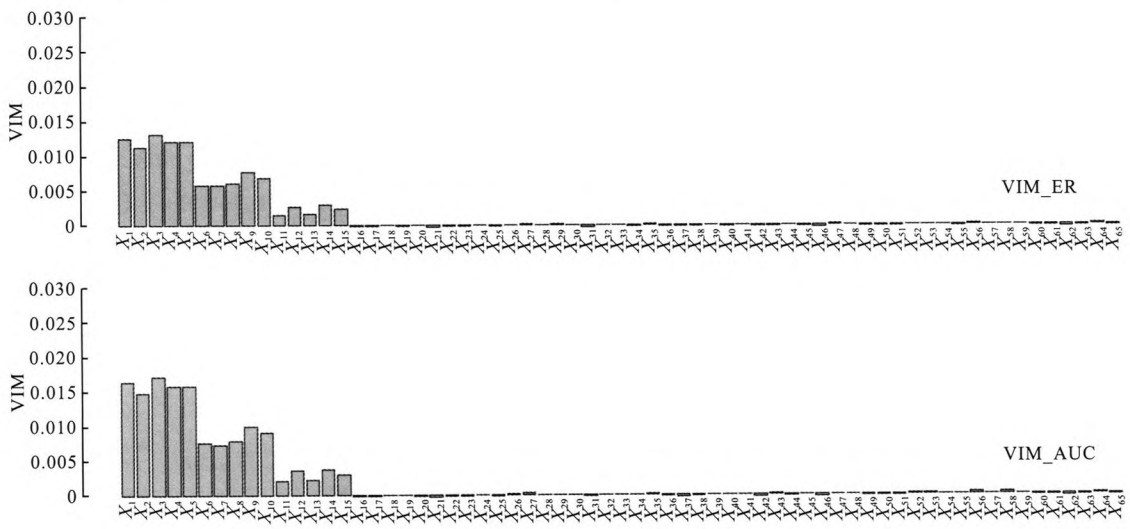

图 1　两组样本量平衡(n_1 = 30，两组样本量比例为 1:1)

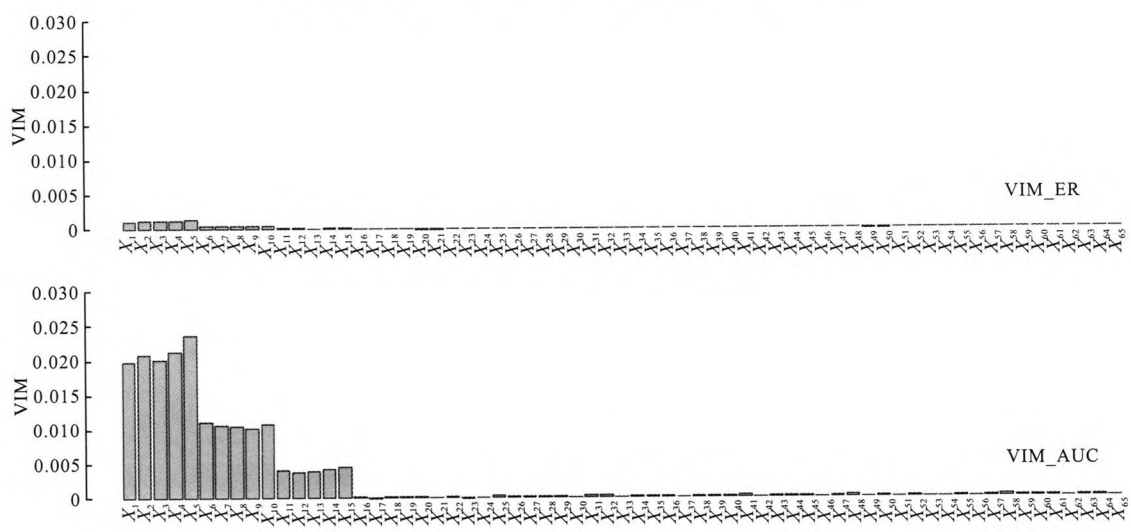

图 2　两组样本量不平衡(n_1 = 30，两组样本量比例为 1:20)

图 4 给出了在不同差异情况下，VIM_ER 和 VIM_AUC 两种方法得到的结果。结果显示，差异不大和样本量较小时，两组不平衡比例对 VIM_ER 的影响非常明显，而 VIM_AUC 则能够更好地区分差异变量与噪音变量。

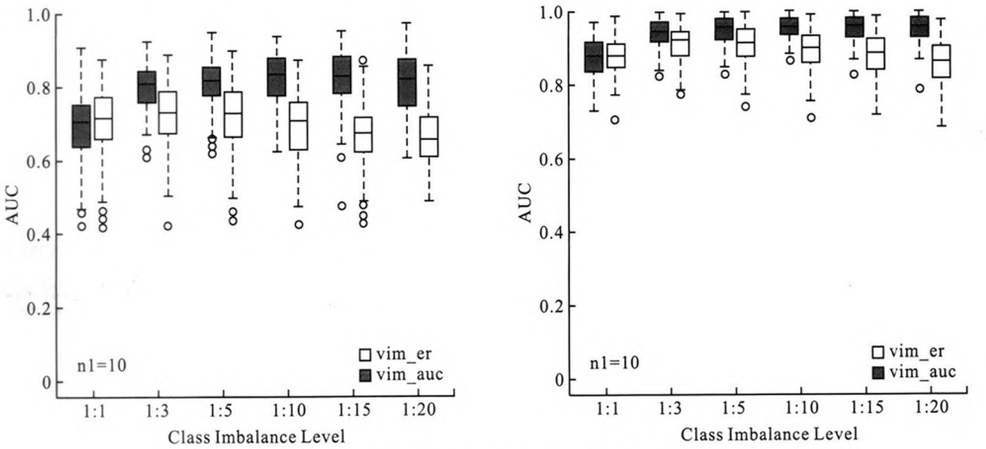

图3 VIM_ER 和 VIM_AUC 两种方法区分 15 个差异变量的能力

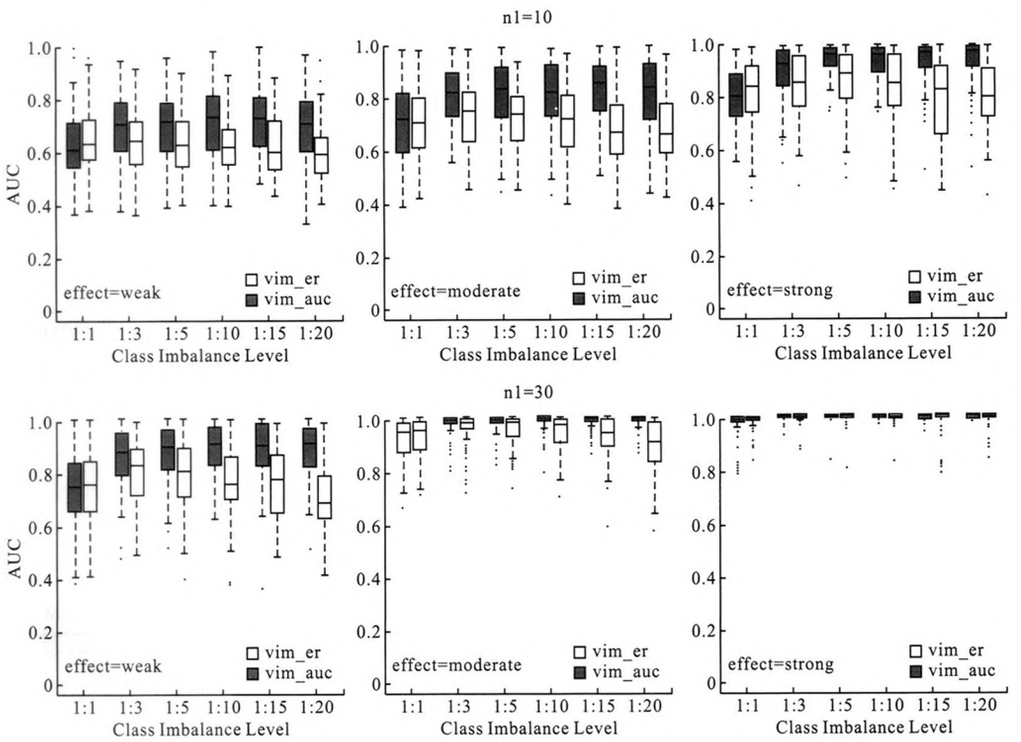

图4 VIM_ER 和 VIM_AUC 两种方法区分 5 个不同差异变量的能力

实际数据验证

本文选取 RNA 编辑数据作为实际数据对上述两种方法进行比较。该数据共包含 2613 例样本，分为两组，其中 1306 例进行了 RNA 编辑，1307 例未进行 RNA 编辑，分析变量 43 个[7]。为评估 VIM_ER 和 VIM_AUC 两种方法在不平衡情况下筛选变量的结果，对数据做以下处理：①随机打乱 43 个变量形成噪音变量，加入到实际数据中，从而共有 43×2=86 个变量；②在第一组中随机抽 100 例，同时在第二组中抽取一定比例的样本，设置两组例数比值分别为 1∶5 和 1∶10。以上过程重复 100 次，最后计算 VIM 得分的平均值。

图 5 分别给出了两组样本量平衡（1∶1）和不平衡（1∶5，1∶10）时，使用 VIM_ER 和 VIM_AUC 两种方法得到的结果。结果显示：两组样本量相同时，VIM_ER 法与 VIM_AUC 法进行变量筛选后得到的 VIM 值排序基本相同；两组样本量不同时，随着两组不平衡程度的增加，使用 VIM_ER 方法得到的 VIM 值中很多逐渐趋于 0，而 VIM_AUC 方法仍能给出相对准确的变量重要性评分，保持“差异变量”的 VIM 值相对较高，从而不会因不平衡问题改变变量的重要性排序。

讨　论

1. 随机森林（RF）是由多个决策树（基分类器）组成的分类器，能够有效地处理非线性、交互作用、共线性以及高维等问题，同时还能够避免过拟合，可以进行预测和变量筛选[8]。在类别间例数不平衡时，实际经常使用的方法是在计算变量重要性时使用错误率，相当于对例数较多的类别赋予了更高的权重，从而导致这种方法估计 VIM 时出现明显的偏倚，这在实际应用

中应予注意。

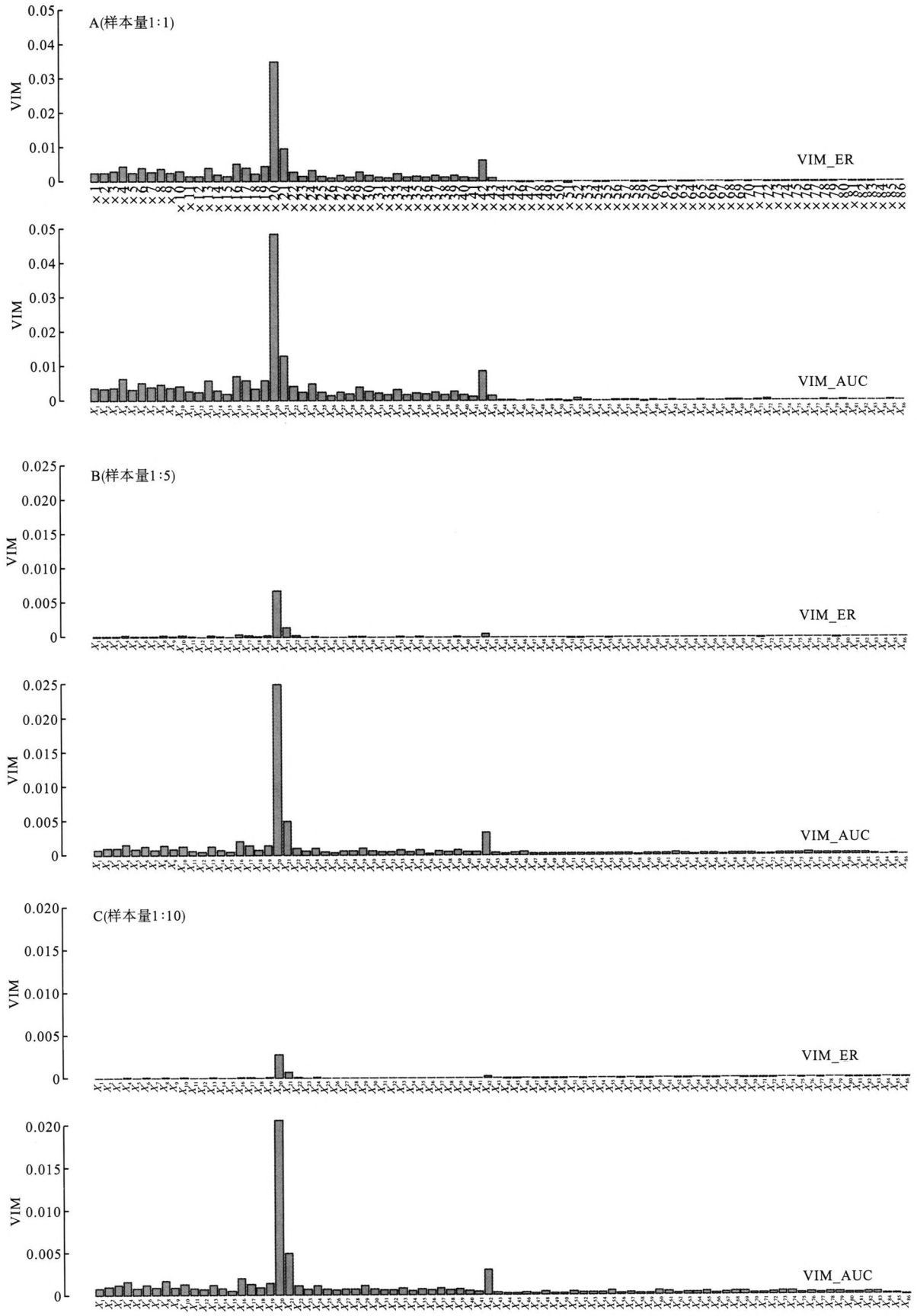

图 5　两种方法的变量重要性评分(A 图 1:1,B 图 1:5,C 图 1:10)

2. 在构建 RF 分类器时,使用 *AUC* 统计量计算 VIM 值,能够在样本例数不平衡时准确地反映变量的作用。模拟实验和实际数据验证的结果显示了这种方法可以有效地解决不平衡的问题。

3. 不平衡的问题主要出现在前瞻性研究中,比如癌症患者远远少于健康人群。这种情况下,虽然可以使用巢式病例-对照的方法,但是如果数据完整,直接

(下转第 542 页)

的、互不相关的、抽象的综合指标，即因子，每个原变量可用这些提取出的公共因子的线性组合表达。为了消除生存质量影响因素间的共线性以及减少影响因素的个数，对生存质量的影响因素进行探索性的因子分析[6-7]，提取公因子、进行因子命名；然后以生存质量得分为因变量，公因子为自变量进行多元逐步回归分析。景睿[20]采用因子分析得出影响农村老年人的生命质量因素，按其作用大小依次为躯体健康因子、心理健康因子、日常生活因子、满意度相关因子、生活条件因子、行为习惯因子和社会关系因子，其中行为习惯因子为负向作用；肖永红[21]应用因子分析得出肺癌患者生存质量很大程度上取决于患者的临床分期、治疗方法、年龄和生存时间。因子分析的优点是通过显在变量测评潜在变量，但其模型和理论尚不完善，没有建立小样本因子分析模型及理论，没有给出因子分析的精确模型等，还有待进一步深入的研究。虽然多种统计分析软件都可实现因子分析，但是 SPSS 窗口化的方式更为方便、简练。目前大多数用因子分析法探讨生存质量影响因素的报道都是用该软件实现的。

小　结

在上面的叙述中，我们介绍了对生存质量影响因素分析的一些统计方法。每一种统计分析方法有各自的适用条件和分析的侧重点。因此，在选择统计分析方法时，要综合考虑研究的目的、资料的特点、分析方法的适用条件、统计分析的侧重点等方面。此外，分析结果要以准确、清晰的形式报告出来，以方便读者对结果的理解。

参 考 文 献

[1] 刘凤斌，方积乾. 医学教育与生存质量. 现代预防医学，2002，29(2)：206-207.

[2] 李振国，杨德森. 生活质量与临床医学. 中国社会医学，1994，3：34-41.

[3] Donnellys. Quality of life assessment in advanced cancer. Current Oncology Reports，2000，29(4)：338-342.

[4] Malmstrom，Klefsqard，Lvarssonb，et al. Quality of life measurements as an indicator for timing of support after oesophagectomy for cancer：a prospective study. BMC Health Serv Res，2015，15(1)：96-99.

[5] 孟欣. 胃癌患者生存质量影响因素分析的方法研究. 硕士论文，2002.

[6] 陈平雁. SPSS13.0 统计软件应用教程. 人民卫生出版社，2005，9.

[7] 张文彤，董伟. SPSS 统计分析高级教程. 高等教育出版社，2013，3.

[8] 徐天和. 中国医学统计百科全书多元统计分析分册. 人民卫生出版社，2004，5.

[9] Schaake，Wiegman，de Groot，et al. The impact of gastrointestinal and genitourinary toxicity on health related quality of life among irradiated prostate cancer patients. Radiother Oncol，2014，110(2)：284-90.

[10] 曲延生，崔哲铭，曲瑶. 大面积烧伤患者长期生命质量调查及影响因素分析. 中国卫生统计，2014，31(6)：1078-1079.

[11] 张娟，陈志远，张志浩，等. 大容量肺灌洗的尘肺患者远期生存质量影响因素分析. 现代预防医学，2010，37(6)：1004-1006.

[12] 隋丹丹，林鹏，陈抒豪. 179 例艾滋病病毒感染者及艾滋病患者生存质量研究. 华南预防医学，2015，41(2)：117-123.

[13] 万丹丹，杨瑞雪，万崇华. 高血压患者生存质量的影响因素分析：QLICD-HY 的应用. 中国卫生统计，2013，30(6)：849-852.

[14] Ely JW，Dawson JD，Mehr DR，et al. Understanding logistic regression analysis through example. Fam Med，1996，28(2)：134-401.

[15] 徐华丽. 广东省城镇居民与健康相关的生存质量状况及其影响因素研究. 硕士论文，2014.

[16] 王玖，胡乃宝，王泽珣，等. 山东省高密市农村老年人生存质量影响因素分析. 中国卫生统计，2014，31(3)：492-493.

[17] 荆春霞，王声湧，吴赤蓬，等. 眼科医务人员生存质量的典型相关分析. 中国公共卫生，2005，21(6)：722-723.

[18] 肖永红，闫子海，张文杰. 肺癌患者生活质量的典型想相关分析. 现代预防医学，2007，34(3)：465-473.

[19] 吴彬，罗佳莹，林声，等. 社会支持与大肠癌患者生存质量的典型相关分析. 中国卫生统计，2014，31(4)：631-635.

[20] 景睿，刘晓冬，李向云，等. 应用因子分析法探讨农村老年人生命质量影响因素. 中国卫生统计，2010，27(3)：309-310.

[21] 肖永红，张翠敏，朱贵东. 因子分析法评价肺癌患者生存质量. 预防医学情报杂志，2010，26(11)：867-869.

（责任编辑：郭海强）

（上接第 540 页）

分析全部数据效果会更好，这时可以使用 VIM_AUC 方法进行变量筛选。

4. VIM_AUC 方法也有一定的局限性，即 *AUC* 这一指标有时不够敏感，因此今后也可以考虑使用部分 ROC 曲线下面积、信息量等其他统计量构建 RF 分类器。

参 考 文 献

[1] Breiman L. Random Forests. Machine Learning，2001. 45(1)：5-32.

[2] Calle M L，Urrea V. Letter to the Editor：Stability of Random Forest importance measures. Briefings in bioinformatics，2011，12(1)：86-89.

[3] Strobl C，Boulesteix AL，Zeileis A，et al. Bias in random forest variable importance measures：Illustrations，sources and a solution. BMC bioinformatics，2007，8(1)：25.

[4] Boulesteix AL，Bender A，Bermejo JL，et al. Random forest Gini importance favours SNPs with large minor allele frequency：impact，sources and recommendations. Briefings in Bioinformatics，2012，13(3)：292-304.

[5] Nicodemus KK. Letter to the editor：on the stability and ranking of predictors from random forest variable importance measures. Briafings in Bioinformatrics，2011，12(4)：369-373.

[6] Janitza S，Strobl C，Boulesteix AL. An AUC-based permutation variable importance measure for random forests. BMC bioinformatics，2013，14(1)：119.

[7] Cummings MP，Myers DS. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. BMC bioinformatics，2004，5(1)：132.

[8] 李贞子，张涛，武晓岩，等. 随机森林回归分析及在代谢调控关系研究中的应用. 中国卫生统计，2012(6)：158-160，163.

（责任编辑：郭海强）