

基于数据挖掘技术的乳腺癌亚型识别方法^①

杨绍华¹, 陈冬东², 张旭¹, 何林¹

1. 西南大学 数学与统计学院, 重庆 400715; 2. 中国科学院 植物研究所, 北京 100049

摘要: 随机森林算法可对特征进行重要性排序, 并能提高运行效率和分类的准确率. 采用方差分析、随机森林算法对乳腺癌基因进行筛选, 使得用随机森林算法、支持向量机算法和 k 近邻算法测试集的准确率分别达到 95.6%, 92.9% 和 92.7%, 并发现了区分乳腺癌不同亚型的两种最重要的基因 GATA3 和 ESR1.

关键词: 数据挖掘; 微阵列; 乳腺癌; 分类

中图分类号: O029

文献标志码: A

文章编号: 1673-9868(2018)05-0113-04

乳腺癌是常见的妇科恶性肿瘤之一, 其诊断、治疗策略和预后往往基于其病理分型. 近年来研究表明, 传统的病理分型对新开展的治疗方法并没有提供有效的指导, 而大量乳腺癌分子生物学特征的数据表明可将乳腺癌分为管腔上皮 A 型(LuminalA 型)、管腔上皮 B 型(LuminalB 型)包括 LuminalB1 型、LuminalB2 型、HER2 过表达型及基底样型(三阴性型、正常乳腺样细胞型)^[1]. 不同的类型其预后以及辅助治疗的效果都有明显差异, 因此识别不同的乳腺癌亚型显得尤为重要.

近年来实验发现, 用免疫组化方法检测药理学标记物(ER, PR, Ki67, HER2)的表达, 可将乳腺癌患者进行亚型分类, 但实验费用昂贵, 实际操作也比较困难, 目前只能局限于实验室, 因此很难广泛应用于临床. 随着计算机技术的发展, 计算机技术不断地应用于生物信息和医疗领域, 取得了显著的成果, 如 Osamu Gotoh 等提出的快速矩阵检测技术可以比较有效地对基因进行检测^[2]. 本文将数据挖掘技术应用到乳腺癌识别中来, 通过建立不同的分类器, 即随机森林、支持向量机和 k 近邻算法, 可以快速地识别乳腺癌不同亚型. 另外随机森林还可以找出乳腺癌不同亚型的高风险基因, 从而对不同亚型的病人实施针对性治疗.

1 数据分析

该数据有 98 个样本, 1 213 个基因, 包含乳腺癌的 3 种亚型(来源于 <http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>). 我们应用分类器对该数据进行分类, 观察经过处理的数据对于 3 种亚型分类的实际符合程度. 这里所采用的分类器性能评价标准是 10 折交叉验证, 且下面得到的准确率均采用 10 折交叉验证得到.

首先用没有经过筛选的数据进行建模, 用随机森林算法得到的准确率仅有 88.0%, 袋外误差(OBB)为 14.0%; 采用支持向量机分类得到的准确率为 83.0%; 采用 k 近邻算法得到的准确率为 90.0%. 这可能是由于基因很多, 而对分类真正有用的却很少, 且若对全部基因进行建模, 计算的工作量及复杂程度将显著增加, 因此需要对基因进行筛选. 本文先采用方差分析进行筛选, 除去表达差异性较小的基因, 然后利用随

① 收稿日期: 2017-06-30

基金项目: 国家自然科学基金项目(11701471); 重庆市基础科学与前沿技术研究项目(cstc2017jcyjAX0476).

作者简介: 杨绍华(1992-), 男, 硕士研究生, 主要从事生物数学研究.

通信作者: 张旭, 副教授.

机森林算法对基因进行重要性排序,再次选择具有重要性的基因,然后对此基因数据进行建模.

1.1 基因的筛选

第一步:方差分析.若一个基因表达水平的分布在目标变量的所有可能值上差异不显著,则可以认为这个基因无助于区分这些目标变量.基于这一点,比较同一类乳腺癌亚型的样本所在数据子集的基因表达水平的均值,也即比较同一目标变量值条件下的基因均值.如果在统计上有很大信心认为一个基因的表达水平均值在属于每一类样本组上没有明显的区别,那么这些基因就可以被排除在进一步分析之外.用方差分析来比较多于两个组的均值,在 R 软件上加载 genefilter 包,可以很方便地进行这种基于方差分析的筛选,其中,假设检验的 p 值设置为 $p<0.01$,筛选后得到 799 个基因.

第二步:用随机森林算法进行进一步筛选.用随机森林算法可得到变量对于分类作用的重要性排序.通过多次交叉验证和袋外误差的比较,所选 32 个基因得到的交叉验证的准确率为 95.6%,袋外误差为 4.08%,所得到的误差最低,分类效果最好.故选出 32 个基因来做进一步分类.

1.2 模型建立及讨论

在筛选数据的基础上分别应用随机森林算法、支持向量机算法和 k 近邻算法进行建模.

首先对数据集设定 10 个交叉验证的数据集,再做随机森林的 10 折交叉验证,这里用到 R 软件程序包 randomFroest 里的 randomforest()函数,其建立决策树棵树的默认值为 500,得到测试集的分类平均准确率为 95.6%.用随机森林算法得到的 32 个基因的重要性如图 1 所示.

其次用支持向量机算法建模.用 R 软件上的程序包 kernlab 中的 ksvm()函数来对乳腺癌进行分类,对于上面设定 10 交叉验证的数据集再做支持向量机的 10 折交叉验证,采用 kernlab 里的 ksvm()函数做预测,得到测试集的分类平均准确率为 92.9%.

同样,再用 k 近邻方法分类,这里所用到的函数是程序包 kknn 的函数 kknn(),通过多次交叉验证, $k=11$ 时为最优邻近数目,虽然 $k=6$ 时,准确率很高,但做交叉验证时,准确率比较低,而不做交叉验证不可能得到最好的信息.因此选择 $k=11$ 对设定的 10 个交叉验证数据集再做最近邻的 10 折交叉验证,得到测试集的分类平均准确率为 92.7%.如图 2,相比没有应用随机森林算法筛选的数据,3 种算法得到的准确率有了有效地提升.

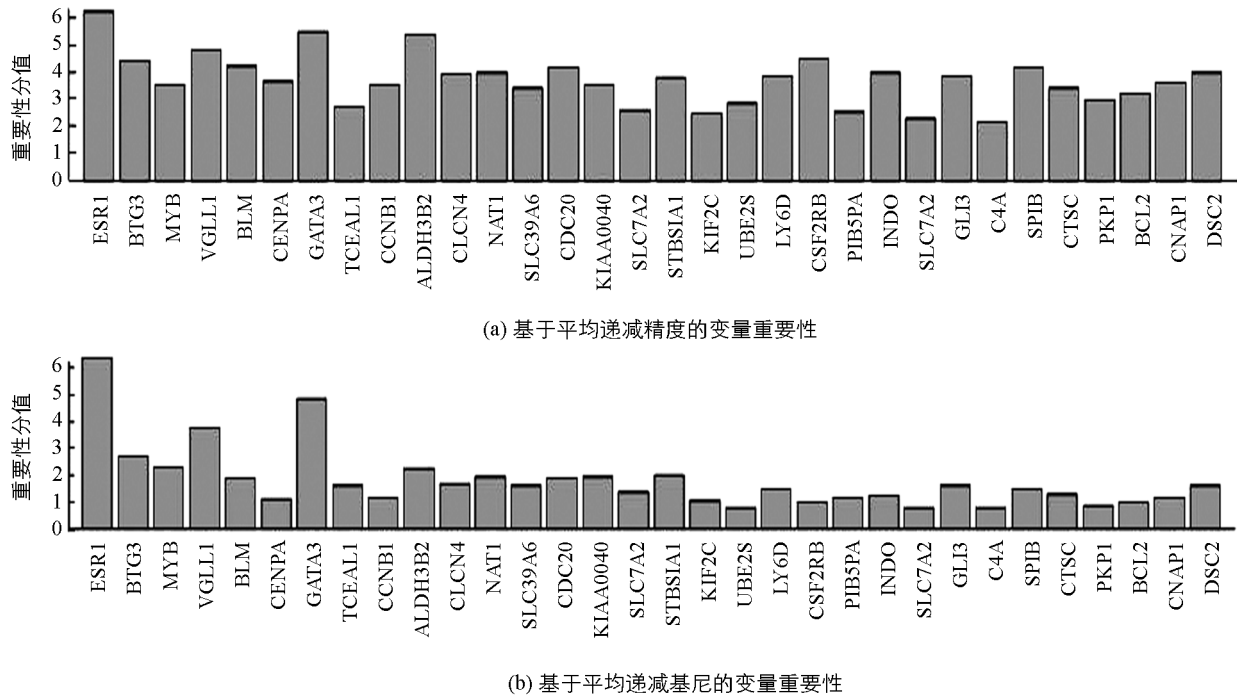


图 1 32 个基因重要性图

由图 1(b)可知, ERS1, GATA3, BTG3 和 VGLL1 等基因的基尼(GINI)指数较高, 这些基因对乳腺癌亚型区分起着十分重要的作用. 相关研究表明, GATA3 是锌指结构转录因子 GATA 家族的一员, 在 T 细胞发育、分化和胸腺发育过程中起重要作用^[3]. GATA3 在乳腺癌中高表达^[4]. 也有报道认为 GATA3 可用于其他肿瘤的鉴别诊断^[5-6]. GATA3 与乳腺癌组织有着密切的关系, 现在许多学者对于 GATA3 在乳腺癌组织中的表达及临床价值进行了相关研究, 发现 GATA3 是在乳腺良性肿瘤和乳腺导管上皮细胞中进行表达, 而在肌上皮细胞中却不表达^[7-9]. GATA3 表达也可能抑制了乳腺癌的发生、发展^[10]. GATA3 不仅与乳腺癌的预后有关, 还可能作为乳腺癌的鉴别诊断标记^[11]. 由此可见, GATA3 对乳腺癌的研究具有十分重要的意义. 对基因 ESR1(雌激素受体基因)编码一种核受体蛋白, 这种蛋白是雌激素敏感组织(如乳腺、子宫内膜和骨骼组织等)对雌激素应答的重要受体. 雌激素通过与受体的高亲和, 发挥其生物学作用, 影响乳腺组织的生长和分化. 该基因的多态性会影响乳腺组织的生长和分化, 进而影响乳腺癌的易感性^[12-14]. 相关研究表明, p53 基因是基因研究最为广泛深入的肿瘤相关基因之一, 是细胞生长周期中负调节因子, 与细胞周期的调控、DNA 修复、细胞分化、细胞凋亡等重要的生物学功能有关, BTG3 作为 p53 的直接靶基因之一, 与人类肿瘤有着生理关联, 是一个候选的肿瘤抑制基因, 因此对 BTG3 进行研究将会促进我们进一步了解肿瘤的发生发展, 对肿瘤的防治有着重要的作用^[15]. 这些基因均对乳腺癌的发生和发展有很大的影响, 这里发现 GATA3, ERS1, BTG3 和 VGLL1 等基因对于乳腺癌不同亚型的区分起到很大的作用.

2 结 论

对于乳腺癌, 在医学上存在着不同亚型的分类, 不同亚型所采取的治疗方法也有很大的差异, 所以识别乳腺癌的不同亚型显得尤为重要. 本文应用方差分析和随机森林算法筛选出对分类有用的基因, 再分别应用随机森林算法、支持向量机算法和 k 近邻算法在测试集上所得准确率分别为 95.6%, 92.9% 和 92.7%, 并发现区分乳腺癌不同亚型的几个最重要的基因 GATA3, ESR1 和 BTG3 等. 这些基因均对乳腺癌的发生和发展有着重要的影响, 同时相比未用随机森林过滤的数据, 准确率得到了很大提升, 为识别不同的乳腺癌亚型提供了帮助, 并且筛选出来的基因远远少于原来的基因, 提高了识别速度, 为诊断乳腺癌高风险基因类别降低了成本.

参考文献:

[1] 孙尚韶, 王玉玺, 梁品. 乳腺癌分子亚型分类及其与新辅助治疗的关系[J]. 中国肿瘤外科杂志, 2011(3): 369—371.

[2] BURSTEIN H J, ELIAS A D, RUGO H S, et al. Phase II Study of Sunitinib Malate, an Oral Multitargeted Tyrosine Kinase Inhibitor, in Patients with Metastatic Breast Cancer Previously Treated with an Anthracycline and a Taxane[J]. Clin Oncol, 2008, 26(11): 1810—1816.

[3] GONZALEZ-ROIBON N, FARAJ S F, MUNARI E, et al. Comprehensive Profile of GATA Binding Protein 3 Immunohistochemical Expression in Primary and Metastatic Renal Neoplasms[J]. Hum Pathol, 2014, 45(2): 244—248.

[4] LI Y, ISHIGURO H, KAWAHARA T, et al. Loss of GATA3 in Bladder Cancer Promotes Cell Migration and Invasion[J]. Cancer Bio & Therapy, 2014, 15(4): 428—435.

[5] MIETTINEN M, MCCUE P A, SARLOMO-RIKALA M, et al. GATA3: a Multispecific but Potentially Useful Marker in Surgical Pathology: a Systematic Analysis of 2500 Epithelial and Nonepithelial Tumors[J]. Am J Surg Pathol, 2014, 38(1): 13—22.

[6] NELSON G. Value of GATA3 Immunostaining in Tumor Diagnosis: a Review[J]. Anat Pathol, 2013, 20(5): 352—360.

[7] 王冬青, 李月峰, 罗一烽, 等. 抑郁症患者杏仁核、海马形态功能变化的 MR 研究[J]. 中华放射学杂志, 2011, 45(7):

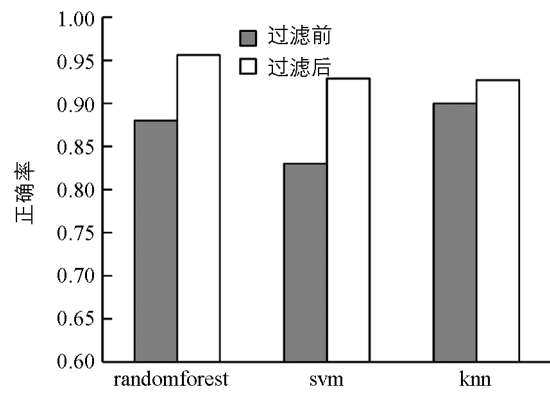


图 2 3 种算法得到的准确率

623—627.

- [8] DOWLATI Y, HERRMANN N, SWARDFAGER W, et al. A Meta-Analysis of Cytokines in Major Depression [J]. *Biol Psychiatry*, 2010, 67(5): 446—457.
- [9] SEDLACIK J, HELM K, RAUSCHER A, et al. Investigations on the Effect of Caffeine on Cerebral Venous Vessel Contrast by Using Susceptibility-Weighted Imaging(SWI) at 1.5, 3 and 7 T [J]. *Neuroimage*, 2008, 40(1): 11—18.
- [10] CLARK B Z, BERIWAL S, DABBS D J, et al. Semiquantitative GATA3 Immunoreactivity in Breast, Bladder, Gynecologic Tract, and Other Cytokeratin 7-Positive Carcinomas [J]. *Am J Clin Pathol*, 2014, 142(1): 64—71.
- [11] 程 凯, 周晓碟, 余 波, 等. 乳腺肿瘤组织中 GATA3 的表达及临床意义 [J]. *临床与实验病理学杂志*, 2015, 31(7): 725—728.
- [12] 张凤春, 徐迎春, 王红霞, 等. 雌激素受体基因 ESR1 多态性与乳腺癌易感性的关系 [J]. *现代肿瘤医学*, 2011, 19(9): 1706—1708.
- [13] DING S L, YU J C, CHEN S T. Diverse Associations Between ESR1 Polymorphism and Breast Cancer Development and Progression [J]. *Clin Cancer Res*, 2010, 16(13): 3473—3484.
- [14] FARMER P, BONNEFOI H, BECETTE V, et al. Identification of Molecular Apocrine Breast Tumours by Microarray Analysis [J]. *Oncogene*, 2005, 24(29): 4660—4671.
- [15] 贺汝燕. 基因在不同人肿瘤细胞中的表达及辐射对其的诱导效应 [D]. 苏州: 苏州大学, 2012.

An Identification Method for Breast Cancer Subtypes Based on Data Mining Technology

YANG Shao-hua¹, CHEN Dong-dong², ZHANG Xu¹, HE lin¹

1. School of Mathematics and Statistics, Southwest University, Chongqing 400715, China;

2. Institute of Botany, Chinese Academy of Sciences, Beijing 100049, China

Abstract: The random forest algorithm can rank features in accordance with their importance and improve the efficiency of operation and the accuracy of classification. In a study reported herein, variance analysis and the random forest algorithm were used to select the characteristics of breast cancer, and the accuracy rate of the random forest algorithm, the CVM (support vector machine) algorithm and the KNN (k -nearest neighbor) algorithm were 95.6%, 92.9% and 92.7%, respectively. Two most important genes, GATA3 and ESR1, were discovered, which can distinguish different subtypes of breast cancer.

Key words: data mining; microarray; breast cancer; classification

责任编辑 张 枸

