两种基于树结构的基因选择算法

谢倩倩[1] 李订芳[1] 章 文[2]
(武汉大学数学与统计学院 武汉 430072)[1] (武汉大学深圳研究院 深圳 518057)[2]

摘 要 癌症诊断是生物信息学领域的重要课题,其中从基因表达数据中选择与癌症相关的基因子集是癌症诊断的关键。随机森林是近年来很热门的算法,它能够评估分类中特征的重要性(该方法简称为PBM)。受此启发,提出了两种基于树结构的基因选择方法FBM和ABM,分别以树结构中特征出现的频率和重要性打分的平均值作为属性重要性的指标。数值实验中,使用提出的方法选取特征子集,并建立随机森林分类器,通过AUC结果评估基因选择的优劣。实验结果表明,当PBM的AUC值不低于0.900时,其在Leukemia数据集上至少需要26个基因,在Colon Cancer数据集上至少需要48个基因。而在仅选取前10个基因时,FBM和ABM在Leukemia数据集的AUC值均达到0.989,在Colon Cancer数据集的AUC值达到0.900。此外,与其它典型的基因选择方法mRMR和ECRP等相比,提出的方法也有较高的精度,这对癌症的精确诊断和及早治疗具有重要的现实意义。

关键词 分类,基因选择,随机森林
中图法分类号 TP3-05 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.053

Two Novel Tree Structure-based Methods for Gene Selection

XIE Qian-qian[1] LI Ding-fang[1] ZHANG Wen[2]
(School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)[1]
(Research Institute of Shenzhen, Wuhan University, Shenzhen 518057, China)[2]

Abstract Cancer diagnosis is one of the most significant topics in bioinformatics. For the microarray datasets, selecting a small subset of genes from thousands of genes (named gene selection) is helpful for accurate identification and treatment of cancerous tumors. Motivated by the instinct of random forests measuring variable importance (named 'PBM'), we proposed two novel methods based on the tree structures for gene selection, namely FBM and ABM. They respectively make use of gene frequency and average scores yielded by a great number of decision trees, which are constructed on the microarray datasets. In computational experiments, the optimal gene subsets are determined by three methods, and random-forest classifiers are built on subsets to evaluate the performance of gene selection methods. AUC scores of PBM are greater than 0.900 when selecting 26 genes for leukemia dataset and 48 genes for colon cancer dataset, while the classifiers with FBM and ABM can achieve the AUC score of 0.989 for leukemia dataset and AUC score of 0.900 for colon cancer dataset respectively with top ten genes selected. In addition, the proposed methods have better performance than the developed methods (such as mRMR and ECRP), which play the critical roles in the accurate diagnosis and treatment of cancer.

Keywords Classification, Gene selection, Random forests

1 引言

随着计算机技术在人类社会各个领域(如社交网络、基因表达微阵列和组合化学等[1,2])的应用,许多实际问题涉及到的数据均具有成千上万维的特征空间,但是实际样本数却很小,这些高维小样本的数据集对传统的机器学习算法提出了一定的挑战。实际的大量研究结果却表明,真正表征事物本质的特征只是其中的很小部分,利用小部分特征进行分类和回归[3]的效果要远优于利用全部特征[4],因此引入了特征选择算法[5,6]。特征选择指的是从高维数据中选取最优特征子集用于分类和回归,它在视频语义检测[7]、文本分类[8]、生物医学诊断[9]及入侵检测[10]等方面均有广泛的应用。其中,在生物医学诊断方面,尤其是癌症诊断上的应用,是信息和决策领域的研究重点。

癌症是一种致命的疾病,每年都会导致上千万人死亡。从生物学的角度讲,基因通常被看作是癌症的诱因。在大量

到稿日期:2014-06-21 返修日期:2014-07-28 本文受国家自然科学基金(61271337,61103126),教育部博士点基金(20100141120049),湖北省自然科学基金(2011CDB454),深圳市战略新兴产业发展专项资金项目(JCYJ20130401160028781)资助。
谢倩倩(1988-),女,硕士生,主要研究方向为机器学习、生物信息;李订芳(1966-),男,博士,教授,主要研究方向为计算流体力学、科学与工程计算软件、计算机应用;章 文(1981-),男,博士,副教授,主要研究方向为数据挖掘、机器学习、生物信息,E-mail:zhangwen@whu.edu.cn(通信作者)。

的人类基因中，只有极少比例的基因对细胞的癌变有着至关重要的作用，因此基因选择作为特征选择的一种应用，得到了越来越多的关注[11]。基因选择是指从微阵列数据中剔除大量冗余或无关基因的过程，从而简化原始数据集，进而提高分类性能。但是基因选择面临的一大挑战是样本数很小（通常小于 100），而基因数很大（通常是数以千计甚至数以万计）[12]。因此，效率和泛化性能是基因选择算法首要考虑的因素[9]。

近年来，随机森林（RF）[13]在生物医学和生物信息学方面得到了越来越多的关注。作为一种集成选择算法，它不仅运行高效，而且对冗余特征不敏感，与其他机器学习方法相比具有更多的优势。此外，RF 本身可以评估基因的重要性，受其启发，本文提出了两种新颖的基于树结构的基因选择方法。本文第 2 节介绍新提出的算法；第 3 节是实验的结果及讨论；最后给出研究结论。

2 方法

2.1 相关研究

基于树结构的方法因其较强的解释性，在机器学习领域（如回归树（CART）和随机森林（Random Forest，RF））备受欢迎。

CART 是由 Breiman 等人在 1984 年提出的[14]。它由分类树和回归树两部分组成，并且采用基尼指数作为分裂标准。Bagging 算法[15]是由 Breiman 于 1996 年提出的，是一种通过操作训练样本集来生成各异的子分类器的方法。RF 是一种组合分类器算法，是树型分类器的集合。它利用 Bagging 算法[15]生成不同的抽样数据集，然后采取 CART 算法建立多个基础分类器。最终的分类或回归结果通过大多数投票法或简单平均法得到。RF 由于集成了 Bagging 和随机选择特征分裂两种方法的特点，因此具有运行效率高、可以评估分类特征或者属性的优势。

用 RF 来评估特征重要性的方法 PBM（permutation-based-method）[13]的基本思想分为 3 部分：（1）在原始数据集上建立 RF，并计算其 10-CV 后的 AUC 值；（2）将原始数据集中的每一个属性对应的所有值进行重排，其他属性值保持不变，并计算 10-CV 后的 AUC 值；（3）计算原始数据集与重排数据集对应的 AUC 的差值，并将该差值作为属性重要性的评价标准。PBM 的流程见图 1。

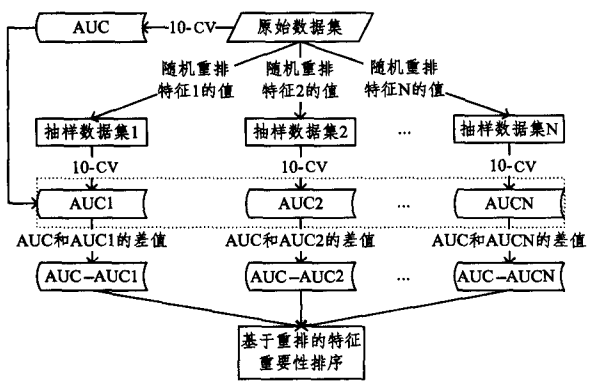


图 1 PBM 的流程

2.2 基于树结构的基因选择方法

受 PBM 的启发，本文提出两种基于树结构的基因选择方法，它们分别以树结构中基因出现的频率和重要性打分的平均值作为评价标准，即 frequency-based-method 和 average-based-method，简记为 FBM 和 ABM。

对于一个包含 M 行 N 列的数据集（$M \ll N$），M 行代表 M 个不同的样本，N 列代表 N 个不同的特征。FBM 是在原始数据集上建立多棵树，将每个属性在所有树中出现的频率作为属性重要性的评价标准。FBM 的基本思想分为 4 步：（1）从原始数据集中随机抽取一个 M 行 n 列的子集（$n \ll N$）；（2）在子集上建立决策树；（3）重复上述步骤建立 k 棵决策树，并记录每个属性在 k 棵树中出现的频率；（4）将属性按照其出现的频率进行排序。出现频率高的属性比频率低的属性更为重要。FBM 的流程见图 2。

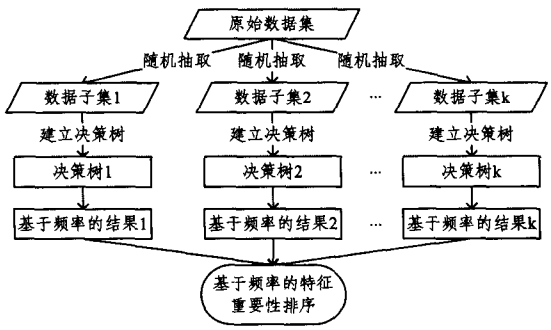


图 2 FBM 的流程

ABM 建立多棵决策树，并将每棵决策树对同一特征的重要性打分的平均值作为特征重要性的评价标准。决策树对每个特征的重要性打分，是指由某一特征作为分裂节点导致的特征重要性的变化值与分支节点个数的比值。如果没有代孕分裂，分支节点数就是所有分支上的分裂数；否则，分支节点数就是每个分支节点包含代孕分裂在内的所有分裂数。ABM 的基本思想也分为 4 步：前两步与 FBM 相同；第 3 步是建立 k 棵树，并且每棵树都对特征打分；最后用 k 棵树的重要性打分的平均值作为最终的评价标准。FBM 的流程见图 3。

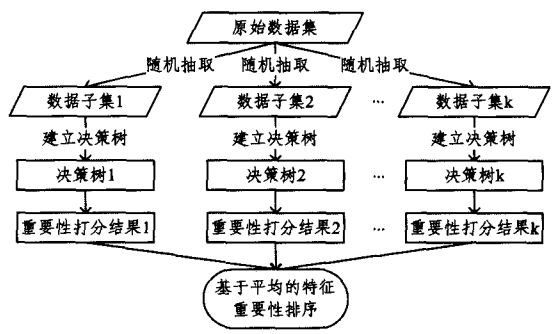


图 3 ABM 的流程

3 实验与分析

3.1 数据集

用两个公开的数据集来评估和比较基因选择算法。这两个公开的数据集分别是 Leukemia 数据集[4]和 Colon Cancer 数据集[16]，都是两分类问题。Leukemia 数据集来源于 Whitehead 研究所和麻省理工学院基因研究中心[4]，它被 Golub 等人首次引用之后，得到了众多研究者的关注。Colon Cancer 数据集曾被 Alon 在文献[16]中进行研究。两个数据

集的详细信息见表 1。

表 1　实验中的两个公开数据集

数据集	来源	基因个数	样本数	类别数	类别名称	各类样本数
Leukemia	Golub 等 (1999)	7129	72	2	ALL	47
					AML	25
Colon Caner	Alon 等 (1999)	2000	62	2	Tumor	40
					Normal	22

3.2　实验设置

无论是 FBM 还是 ABM，包含 M 个样本 $n(n \ll N)$ 个特征的子集都是从原始数据集中随机抽取的。实验中，设定 n 为不超过 N 的平方根的最大整数，原因在于随机森林中每个节点的分裂特征数默认为特征总数的平方根。此外，k 取值为 500，原因也在于随机森林中默认值为 500。

文中采用了 10-CV，将数据集随机划分为 10 份，每次以其中的 9 份作为训练集，余下的 1 份作为测试集。数值实验中，对于本文采用的 Leukemia 和 Colon Cancer 这两个数据集，均在其特征子集上建立随机森林的分类器，以分类器的好坏作为分类评价的标准，而不是以唯一确定的基因数目作为评价的标准。

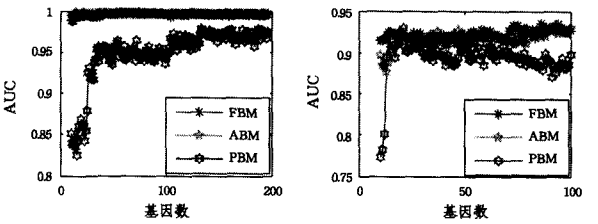
此外，本文采用 AUC 值作为分类器好坏的评判标准。通过不断变化分类的阈值，可以得到一条表示错分为正类样本的比重与错分为负类样本的比重之间的变化曲线（即 ROC 曲线），比重越小表明分类器的分类性能越好。ROC 曲线虽然很好地反映了分类器的性能表现，但是无法准确地比较两条 ROC 曲线，因此选择采用 ROC 曲线下的面积（即 AUC）将其量化，AUC 值越大表明分类性能越好。

3.3　与随机森林自带的方法比较

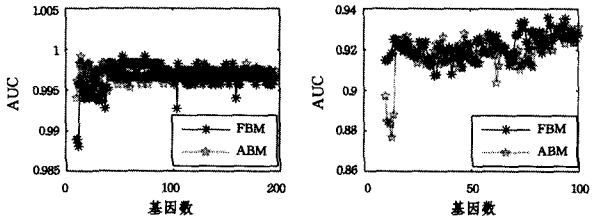
为了分析提出的两种方法（FBM 和 ABM）与随机森林自带的方法（PBM），对 3 种方法分别选取相同数目的基因子集，然后在这些子集上建立随机森林分类器，以分类器的性能作为 3 种基因选择方法的评估标准。

将这 3 种基因选择方法分别应用于 Leukemia 数据集和 Colon Cancer 数据集。对于所有的方法和所有数据集，仅有极少部分基因具有很高的重要性打分，且 FBM 和 ABM 两种方法的基因重要性排序比 PBM 明显很多。因此，FBM 和 ABM 均比 PBM 有更好的分类性能。

进一步讨论 3 种方法在各个基因子集上的分类性能。先选取包含前 10 个基因的子集，依次递增 1 个基因构成新的基因子集。对于 Leukemia 数据集，最大的基因子集包含 200 个基因，而 Colon Cancer 数据集包含 100 个基因，它们对应的 AUC 值的变化见图 4。从图 4(a)和图 4(b)可以看出，对于所有的方法，AUC 值都是随着基因数的增加而有所增加的，但是到达一个峰值以后有所下降。这一现象表明基因数越多不一定导致更好的分类性能，而且比较耗时。总体而言，3 种方法都能产生较好的分类性能，但是 FBM 和 ABM 的表现优于 PBM。尽管 PBM 是 3 种方法中最差的，但是在 Leukemia 数据集上选取 26 个基因和 Colon Cancer 数据集上选取 48 个基因时，其 AUC 值也均达到了 0.904。在仅选择十几个基因的情况下，PBM 在 Leukemia 数据集上的 AUC 值达到 0.851，在 Colon Cancer 数据集上的 AUC 值达到 0.905；同等条件下，FBM 和 ABM 在 Leukemia 数据集和 Colon Cancer 数据集上的 AUC 值均在 0.995 和 0.915 左右。此外，FBM 和 ABM 所能达到的最好的 AUC 值在 0.998 左右。鉴于 FBM 和 ABM 的分类性能极为相似，接下来的实验重点考察这两类方法的优劣性。从图 4(c)和图 4(d)可以看出，FBM 和 ABM 在 Leukemia 数据集上均具有很好的表现。在图 4(c)中，FBM 和 ABM 在仅选取 10 个基因时，它们的 AUC 值就达到了 0.989，并且当基因数目在 55 和 73 之间变动时，FBM 的 AUC 值一直维持在 0.998 左右，但是 ABM 的 AUC 值仅为 0.996。因此，在 Leukemia 数据集上，FBM 的性能略优于 ABM。同样，对于数据集 Colon Cancer，为了取得 0.915 的 AUC 值，FBM 仅需要 10 个基因，而 ABM 却需要 15 个基因。因此，在 Colon Cancer 这个数据集上，FBM 的表现也优于 ABM。



(a)3 种方法在 Leukemia 数据集上的对比结果　(b)3 种方法在 Colon Cancer 数据集上的对比结果



(c)2 种方法在 Leukemia 数据集上的对比结果　(d)2 种方法在 Colon Cancer 数据集上的对比结果

图 4　在两个公开数据集上的对比结果

总体而言，本文提出的两种方法均比 PBM 有更好的表现，不仅易于实施，而且易于解释。PBM 独立地评估基因的重要性，没有考虑到基因间的关联性。而 FBM 和 ABM 通过树结构评估基因的重要性，在建树的过程中综合考虑了基因间的关联性，以所有树的统计特征作为最终的评判。此外，FBM 和 ABM 均比 PBM 高效，而且分类性能也优于 PBM。

3.4　与其他基因选择算法的比较

为了进一步阐述本文出提的两种基因选择方法的优越性，将 FBM 和 ABM 分别与 mRMR[17]、ECRP[18]、RBF[19]、Relief[19]、ACA[20]等方法进行比较，它们都是最具代表性的基因选择方法。mRMR[17]和 ECRP[18]分别是由 Chris Ding 等人和 Hojin Moon 等人提出的。其中，mRMR 选择与类标签有较大相关度而与已选择的基因有较小冗余度的基因；ECRP 则是一种集成基因选择法，它从原始空间中随机抽取大量基因子集，然后在这些基因子集上建立分类器，并通过大多数投票法进行整合。

在下面的实验中，为了公平全面起见，对于 FBM、ABM、mRMR[17]和 ECRP[18]中的每种基因选择方法都选择相同数量的基因；同时，在已选定的基因子集上，分别采用贝叶斯算法（Naïve Bayes，NBC）和 k 近邻算法（k-nearest-neighbor，kNN）来建立分类器。实验中，kNN 中的 k 取值为 3，简记为 kNN。此外，采用最受欢迎的交叉验证来评估各种基因选择法的表现，实验中分别验证了五折交叉验证（5-fold cross-validation，5-fold）和十折交叉验证（10-fold cross-validation，10-fold）。

表2概括了4种不同的基因选择法以kNN为分类器时的分类误差，并分别采用5-fold和10-fold作为评估的标准。从表中可以看出，无论是以5-fold为评估标准还是以10-fold为评估标准，FBM和ABM两种基因选择方法在Colon Cancer数据集上的表现均优于其他两种方法；而在Leukemia数据集上，虽然FBM和ABM的表现次于mRMR，但是远优于ECRP。对于以NBC为分类器的情况，可得出同样的结论。

表2 4种基因选择法在两个数据集上5-fold和10-fold验证后的分类误差(kNN为分类器)

数据集	5-fold				10-fold			
	FBM	ABM	mRMR	ECRP	FBM	ABM	mRMR	ECRP
Colon	12.9	12.9	22.58	29.03	12.9	12.9	24.19	30.65
Leukemia	4.17	2.78	0.00	16.67	2.78	2.78	0.00	15.28

在文献[19]中，以NBC作为分类器的RBF选择法在数据集Colon Cancer和Leukemia上的分类准确率分别为88.71%和98.61%(其所选择的基因个数分别为4和16)，同时Relief相应的准确率分别为85.48%和97.22%。FBM在数据集Colon Cancer和Leukemia上分别得到了72.58%和97.22%的准确率，ABM得到了88.71%和95.83%的准确率。Au等人[20]提出的基于属性聚类的基因选择算法(ACA)也以NBC为分类器，在数据集Colon Cancer和Leukemia上分别选取前7个和前50个基因时，其分类误差分别为35.5%和38.2%，均高于本文所提方法的误差：FAM的误差分别为17.74%和4.17%，ABM的分类误差分别为12.90%和4.17%。Yang等人[21]采用kNN作为分类器，在数据集Leukemia选取前30个基因时，其分类误差为5.6%，而FAM和ABM的分类误差均为5.56%。此外，以kNN作为分类器时，FAM和ABM在两个数据集上的表现也均优于ACA。

结束语 受到随机森林自带的特征重要性评估方法的启发，本文提出了两种基于树结构的方法应用于癌症诊断中的基因选择问题。在两个公开数据集上的检验结果表明，提出的方法优于随机森林自带的特征重要性评估方法；与重要的基因选择方法mRMR、ECRP、RBF、Relief和ACA等相比，提出的算法也具有优势。因此，本文提出的新算法对于选择癌症相关基因、实现癌症诊断具有一定的实际意义。

参考文献

[1] Xing E P, Jordan M I, Karp R M. Feature selection for high-dimensional genomic microarray data[C]// Proceedings of the 15th International Conference on Machine Learning. 2001:601-608

[2] Andrew Y N. On feature selection: learning with exponentially many irrelevant features as training examples[C]// Proceedings of the 15th International Conference on Machine Learning. 1998:404-412

[3] Bhattacharjee A, Richards W G, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses [J]. Proceedings of the National Academy of Sciences of the United States of America, 2001,98(24):13790-13795

[4] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer, class discovery and class prediction by gene expression monitoring[J]. Science, 1999,286(5439):531-537

[5] Faivishevsky L, Goldberger J. Unsupervised feature selection based on non-parametric mutual information [C]// 2012 IEEE International Workshop on Machine Learning for Signal Proceeding (MLSP). IEEE, 2012:1-6

[6] 冶晓隆，兰巨龙，郭通. 基于PCA和禁忌搜索的网络流量特征选择算法[J]. 计算机科学，2014,41(1):187-191
Ye Xiao-long, Lan Ju-long, Guo Tong. Algorithm of Network Traffic Feature Selection Based on PCA and Tabu Search[J]. Computer Science, 2014,41(1):187-191

[7] Zhu Qiu-sha, Lin Lin, Shyu Mei-ling, el al. Feature Selection Using Correlation and Reliability Based Scoring Metric for Video Semantic Detection[C]// IEEE Fourth International Conference on Semantic Computing. 2010:462-469

[8] Ogura H, Amano H, Kondo M. Comparison of metrics for feature selection in imbalanced text classification [J]. Expert Systems with Applications, 2011,38(5):4978-4989

[9] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007,23(19):2507-2517

[10] Amiri F, Yousefi M R, Lucas C, et al. Mutual information-based feature selection for intrusion detection systems [J]. Journal of Network and Computer Applications, 2011,34(4):1184-1199

[11] 于化龙，顾国昌，赵靖，等. 基于DNA微阵列数据的癌症分类问题研究进展[J]. 计算机科学，2010,37(10):16-32
Yu Hua-long, Gu Guo-chang, Zhao Jing, et al. State of the Art on Cancer Classification Problems Based on DNA Microarray Data[J]. Computer Science, 2010,37(10):16-32

[12] Liu Jing-jing, Cai Wen-sheng, Shao Xue-guang. Cancer classification based on microarray gene expression data using a principal component accumulation method [J]. Science China Chemistry, 2011,54(5):802-803

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

[14] Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees [M]. Chapman and Hall/CRC, 1984

[15] Breiman L. Bagging predictors [J]. Machine Learning, 1996,24(2):123-140

[16] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Sciences of the United States of America, 1999,96(12):6745-6750

[17] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data [J]. J Bioinform Comput Biol, 2005,3(2):185-205

[18] Moon H, Ahn H, Kodell R L, et al. Ensemble methods for classification of patients for personalized medicine with high-dimensional data [J]. Artif Intell Med, 2007,41(3):197-207

[19] Yu L. Feature selection for genomic data analysis[M]// Computational methods of feature selection. Chapman & Hall, 2008:337-353

[20] Au W-H, Chan K C C, Wong A K C, et al. Attribute clustering for grouping, selection, and classification of gene expression data [J]. IEEE/ACM Trans Computational Biology and Bioinformatics, 2005,2(2):83-101

[21] Yang Kun, Cai Zhi-peng, Li Jian-zhong, et al. A stable gene selection in microarray data analysis [J]. BMC Bioinformatics, 2006,7:228