文章编号：2095-2163(2019)06-0021-03 中图分类号：TP391 文献标志码：A

基于生物学通路的癌症分类研究

张巧生[1,2]，李 杰[1]
（1 哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001；2 黑龙江八一农垦大学 理学院，大庆 163319）

摘 要：尽管基因标志物已广泛成功应用，但是仍存在很多问题。其一是在疾病的发展和治疗反应中识别出的很多基因标志物缺乏合理的生物学功能解释，其二是针对癌症这种异质性疾病，基因标志物的可重复性是一大挑战。基于此，本文提出了一个以生物学通路为特征的分类方法。实验结果表明该方法在分类性能上优于基于以基因为特征的分类算法。

关键词：生物学通路；分类；癌症

Cancer classification research based on biological pathway

ZHANG Qiaosheng[1,2], LI Jie[1]
(1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China;
2 College of Science, Heilongjiang Bayi Agricultural University, Daqing, 163319)

【Abstract】 Despite the success of gene biomarkers' s use, they have not been exempt of problems. Specifically, one major drawback of multi-gene biomarkers is that they often lack proper interpretation in terms of mechanistic link to the fundamental cell processes responsible for disease progression or therapeutic response. The other major drawback is that these gene signatures have been challenging to reproduce, particularly in heterogeneous diseases such as cancer. Accordingly, this paper proposes a classification method based on biological pathways. The experimental results show that the proposed method outperforms the gene-based classification algorithm.

【Key words】 biological pathway; classification; cancer

0 引 言

随着用于全基因组表达谱分析的高通量技术的出现，研究人员提出了许多方法来发现癌症相关驱动基因作为指导癌症诊断和预后的生物标志物[1-4]。然而，对于癌症这种高度异质性疾病，这些基因特征往往是不具有可重复性的。此外，诸如噪声、测量误差和大量的基因假说等等因素也会阻碍实验结果的可重复性。同时，发现的这些癌症相关驱动基因列表与疾病进展或治疗反应相关的生物学过程往往很难建立联系，生物学意义不清晰。随着研究的深入，人们越来越认识到基于通路的分析可以克服上述缺陷。通过将基因水平数据折叠成紧凑、功能性的通路水平数据，不但可以压缩特征，还可以减少过拟合，提高概括性，同时保持生物可解释性[5]。

1 算法描述

本文提出了一种基于生物学通路的癌症分类方法。首先通过 Pathifier 算法[6]把基因水平数据转换成通路水平数据，然后基于相关特征选择（correlation feature selection，CFS）进行特征选择，最后基于选择后的特征使用 SVM 分类模型在测试集上进行分类效果评价。为了验证本文方法的有效性，文中方法与基于基因生物学标记的癌症分类方法[7]进行了比较分析。

1.1 Pathifier 算法

Pathifier 算法通过单个癌症样本下通路对所有控制样本下通路中值的偏离程度来计算单个癌症样本下的通路分数。下面详细描述 Pathifier 算法原理。

假设给定通路基因列表 $K(|K|\geqslant 3)$。基因表达数据根据通路基因列表构建 $|K|$ 维空间，每个基因代表一个维度，空间中的每个点代表一个样本。所有的样本点构成 $|K|$ 维空间中的点云，设样本点个数为 n。然后根据 Hastie and Stuetzle 算法[8]在点云中寻找主曲线 $f(\lambda)$，其中 λ 为主曲线的点，如图 1A 所示，不同颜色的点代表不同表型下的样本。假设 x 为空间中的点，其对应的 λ 由公式(1)求得。

基金项目：国家自然科学基金(61471147)。

作者简介：张巧生(1981-)，男，博士研究生，讲师，主要研究方向：生物信息学、数据挖掘；李 杰(1971-)，男，博士，副教授，博士生导师，主要研究方向：大数据分析和挖掘、模式识别、生物信息学。

收稿日期：2018-12-24

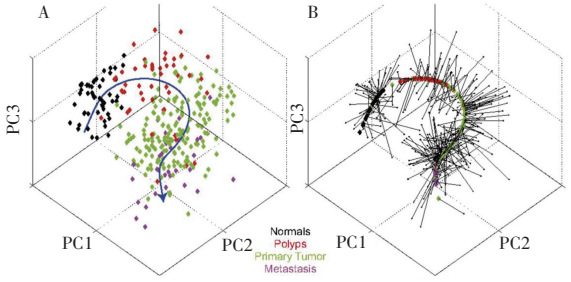

图 1 细胞凋亡通路在结肠癌数据集中的主曲线[6]

Fig. 1 The principal curve learned for the apoptosis pathway on the colorectal dataset[6]

$$f(\lambda) = \mathbb{E}(X \mid \lambda_f(X) = \lambda). \tag{1}$$

其中，$\lambda_f(x) = \sup_{\lambda}\{\lambda: \|x - f(x)\| = \inf_{\mu}\|x - f(\mu)\|\}$，$X \in M_{n\times|K|}(\mathbb{R})$。

找到主曲线 $f(\lambda)$ 后，样本 x 投影到曲线 $f(\lambda)$ 上最近的点就代表该样本在主曲线上位置，如图 1B 所示。由部分正常(Normal)样本形成的质心为主曲线的起始点，如图 1A 所示。则每个样本下的通路分数就等于该样本在主曲线上的位置沿曲线到起始点的距离。

基于 Pathifier 算法，基因水平数据就可以转换成通路水平数据。

1.2 特征选择

通过把基因表达值转换为通路水平得分，基因表达矩阵转化为通路得分矩阵。为了优化预测模型，本文采用一种基于关联的特征选择(Correlation based Feature Selection，CFS)方法。CFS 是一种过滤型(Filter)特征选择算法[9]，其启发式的筛选与表型高度相关且彼此不相关的特征子集作为预测特征。不相关的特征被忽略，因为其与表型具有很低的相关性。冗余特征应被剔除，因为其与一个或多个剩余特征高度相关。

1.3 算法评价

本文选用 SVM 算法构建分类模型。数据集根据表型分层随机抽样分成 3 部分，三分之二用于特征选择和训练，三分之一用于测试。评价指标为评价分类性能的常用指标，分别为准确率(*Accuracy*)、召回率(*Recall*)、精确率(*Precision*)、*F* 值(*F - score*)、*ROC* 曲线下面积(*AUC*)。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$Recall = \frac{TP}{TP + TN}, \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$F - score = \frac{2 \times TP}{2TP + FP + FN}. \tag{5}$$

其中，*TP*(True Positive)即真阳性，是指属于类别 C 被分类成类别 C 的样本个数、*TN*(True Negative)即真阴性，是指非类别 C 而被分成非类别 C 的样本个数；*FP*(False Positive)即假阳性，是指非类别 C 被分成类别 C 的样本个数；*FN*(False Negative)即假阴性，是指属于类别 C 而被分成非类别 C 的样本个数。

2 数据集

实验中的验证数据集(ID= GSE25066)下载自 GEO。GSE25066 数据集共包含 488 个样本，其中病理完全缓解(pathologic complete response，PCR)样本 99 个，残留病灶(residual disease，RD)样本 389 个。这个数据集是乳腺癌关于新辅助化疗(neoadjuvant chemotherapy，NAC)效果的数据集。PCR 样本通过新辅助化疗达到病理完全缓解的患者，RD 样本是对新辅助化疗不敏感的患者。研究表明通过新辅助化疗达到 PCR 的患者，无病生存(disease free survival，DFS)以及总生存(overall survival，OS)均得到显著的提高。

实验中通路数据来自 KEGG(Kyoto Encyclopedia of Genes and Genomes)通路数据库(PATHWAY database)[10]。1995 年，日本京都大学生物信息学中心的 Kanehisa 实验室人工构建了 KEGG 数据库，是基于使用一种可计算的形式捕捉和组织实验得到的知识而形成的系统功能知识库。KEGG 通路基因集下载自 ConsensusPathDB 网站(http://consensuspathdb.org/)。经过筛选，选出 281 个 Homo sapiens(hsa)通路作为本文实验用通路数据。

3 实验结果与结论分析

Pathifier 算法在基因表达矩阵转化为通路得分矩阵时，由于有 3 个通路包含的基因个数少于 3 个，所以实际应用中只有 278 个通路成功转化。基因表达数据中，PCR 样本往往比 RD 少很多，存在着类别不平衡现象。为了消除类别不平衡对基分类器的影响，在类别多的 RD 样本中随机抽取 PCR 样本个数的 RD 样本来平衡数据集，即从 RD 样本中随机抽取 99 个样本。通过 CFS 算法最终筛选出 32 个特征用于训练模型，最后在测试集上评估算法性能。为了验证方法的有效性，本文所提方法与文献[7]中基于基因特征方法进行了比较分析，实验结果如图 2、

图 3 所示。本文基于通路(pathway)方法的准确率、召回率、精确率、*F* 值和 *AUC* 分别为 65.15%，78.78%，61.90%，69.33%，69.74%。由图 2、图 3 可以看出，本文方法整体性能要优于基于基因的方法。

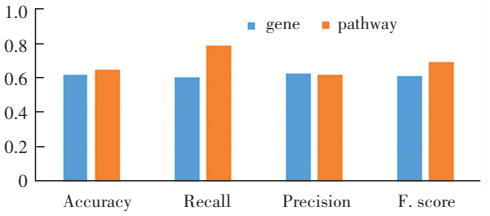

图 2　两种方法的性能比较

Fig. 2　Performance comparison between the two methods

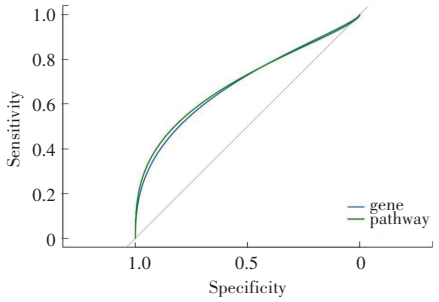

图 3　两种方法的 ROC 曲线

Fig. 3　ROC curve for two methods

4　结束语

针对以基因为特征分类算法的特征不可重复性和相关特征的生物学意义不明确，本文提出了一个以通路为组学特征，结合相关特征选择(CFS)和分类算法预测乳腺癌用药反应的方法。实验结果表明，本文方法的分类性能优于基于基因为特征的分类算法，而且生物学通路的生物学意义明确，为在功能机制上深入了解癌症致病机理提供了新的途径。

参考文献

[1] PAIK S, SHAK S, TANG G, et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer[J]. New England Journal of Medicine, 2004, 351(27):2817-2826.

[2] SOTIRIOU C, PUSZTAI L . Gene-expression signatures in breast cancer.[J]. New England Journal of Medicine, 2009, 360(8):790.

[3] VAN " V L J, DAI H, VAN D V M J, et al. Gene expression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871):530-536.

[4] WANG Y, KLIJN J G, ZHANG Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer[J]. Lancet (North American Edition), 2005, 365(9460):0-679.

[5] HAN L, MACIEJEWSKI M, BROCKEL C, et al. A PRObabilistic Pathway Score (PROPS) for Classification with Applications to Inflammatory Bowel Disease[J]. Bioinformatics, 2017.

[6] DRIER Y, SHEFFER M, DOMANY E. Pathway-based personalized analysis of cancer.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(16):6388-6393.

[7] BECKER N. Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data[J]. BMC Bioinformatics, 2011, 12.

[8] HASTIE T, STUETZLE W. Principal curves[J]. Journal of the American Statistical Association, 1989, 84(406): 502-516.

[9] HALL M A. Correlation-based feature selection for machine learning[J], 1999.

[10] KANEHISA M, FURUMICHI M, TANABE M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic acids research, 2016, 45(D1): D353-D361.

(上接第 20 页)

[5] YANG C G, LI J D, NI Q, et al. Interference-aware energy efficiency maximization in 5G ultra-dense networks. IEEE Trans Commun, 2017, 65(2):728

[6] MEMISOGLU E , BASAR E , ARSLAN H. Fading-aligned OFDM with index modulation for mMTC services[J]. Physical Communication, 2019.

[7] List Decoding of Polar Codes [J]. IEEE Transactions on Information Theory. 2015,61(5): 2213-226.

[8] 刘星. 极化码的译码算法研究及实现[D]. 南京大学, 2015.

[9] 尤肖虎,潘志文,高西奇,等.《5G 移动通信发展趋势与若干关键技术》中国科学:信息科学,2014:44(5)

[10] 李世超.5G 关键技术之 NOMA 介绍[J]. 电子制作,2015(4): 139-140.

[11] HAYKIN S. Communication Systems. 4E, ch. 4, pp. 247-308, Wiley 2001.

[12] " Study on LTE device to device proximity services; Radio aspects (Release 12)," TR 36.843 v.12.0.1, 3GPP standardization, Sophia Antipolis, France, Mar. 2014.

[13] STOYAN D, KENDALL W S, MECKE J. Stochastic Geometry and its Applications[J]. Journal of the Royal Statistical Society, 2013, 45(2):345.

[14] 时锐, 杨孝宗. 自组网 Random Direction 移动模型点空间概率分布的研究[J]. 计算机研究与发展, 2004, 41(7):2056-2062.

[15] ELSAWY H, HOSSAIN E, ALOUINI M S. Analytical modeling of mode selection and power control for underlay D2D communication in cellular networks[J]. IEEE Trans. Commun., 2014, 62(11):4147-4161.

[16] 张沛泽, 庞帅, 周宇,等. 5G 高频段无线信道测量技术研究进展及发展趋势[J]. 移动通信, 2017,41(18):67-72.

[17] CHEN K, NIU K, LIN J R. Improved Successive Cancellation Decoding of Polar Codes [J]. IEEE Transactions on Communications, 2013, 61(8): 3100-3107.

[18] VANGALA H, HONG Y , VITERBO E. Efficient Algorithms for Systematic Polar Encoding. IEEE Communications Letters, 2016, 20(1):17-20.

[19] KUMAR D, ASERI T C, PATEL R B. Distributed Cluster Head Election (DCHE) Scheme for Improving Lifetime of Heterogeneous Sensor Networks. Tamkang Journal of Science and Engineering, 2010,13(3): 337-348.

[20] 周坤. 联合信源信道编码方法的研究[D]. 大连理工大学, 2006.