基于主曲线的微阵列数据分类

祁云嵩[1,2] 孙怀江[2]

(江苏科技大学计算机学院 镇江 212003)[1] (南京理工大学计算机学院 南京 210094)[2]

摘 要 提出了一种基于主曲线(principal curves)的微阵列数据分类方法(PC)。主曲线是第一主成分的非线性推广,它是数据集合的"骨架",数据集合是主曲线的"云"。基于主曲线的微阵列数据分类方法,首先利用专门设计的算法在训练数据集上计算出每类样本的主曲线,然后根据测试样本与各类样本主曲线距离的期望方差来确定测试样本所属的类别。实验结果表明,该分类方法在进行小样本微阵列数据分类时性能优于现有的方法。

关键词 基因微阵列,主曲线,模式分类

中图法分类号 TP312 **文献标识码** A

Microarray Data Classification Based on Principal Curves

QI Yun-song[1,2] SUN Huai-jiang[2]

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)[1]

(School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)[2]

Abstract In this paper, a novel classifier was proposed to classify microarray data using principal curves. Principal curves are the non-linear generalization of principal components. Intuitively, a principal curve 'passes through the middle of the data cloud'. As a kind of new classification technique, Principal Curve-based classifier (PC) involves a novel way of computing a principal curve for each class using the training data. A test sample is the class-label of the principal curve that is closest to it according to Expected Squared Error. Experimental results illustrate the performance of the PC is better than other existing approaches when a very small sample size of a microarray set is concerned.

Keywords Microarray data, Principal curve, Pattern classification

1 前言

微阵列技术的应用使得研究人员可以大规模并行提取DNA或RNA信息,从而能够在基因组水平上以系统的、全局的观念去研究生命现象及其本质,包括研究生命体系中不同部位、不同生长发育阶段的基因表达,比较不同个体或物种之间的基因表达,比较正常和疾病状态下基因及其表达的差异,以便于疾病的诊断、分类及治疗等。通过微阵列实验,得到的是包含成千上万个基因的表达数据——基因表达谱,其形式化为基因表达矩阵,行表示基因,列表示实验样本。基因表达数据通常具有数据量大、维数高、样本小、非线性等特点,这对其数据分析方法提出了挑战[1-3]。目前,支撑向量机分类法(SVM)[4]、K近邻分类法(KNN)[5]、贝叶斯分类法(NB)[6]以及C4.5决策树[7]等机器学习方法已普遍用于解决基因表达数据的分类问题,但由于样本采集困难造成的小样本问题,传统的机器学习方法显现出很大的局限性。在像微阵列这样的小样本数据处理问题上,很多成功的分类方法均或多或少地存在"过拟合"问题[8]。

本文提出了一种基于主曲线的分类方法用于微阵列数据的分类。该方法首先利用专门设计的算法计算出训练数据集上每类样本的主曲线,然后根据测试样本与各类样本主曲线的期望距离方差来确定该测试样本的类别。在几个典型的微阵列数据集上的测试结果表明,该方法在一般数据集上的性能仅次于SVM,但在处理极小样本数据集时,本文的分类方法性能要优于现有广泛用于微阵列数据处理的几种分类器。

2 主曲线

主曲线是第一主成份的非线性推广,它对数据分布的结构信息保持性好。根据给定的数据集求出的主曲线是通过数据分布"中间",并满足"自相合"性质的光滑曲线[9,10]。所谓自相合,即曲线上的每一点都是投影至该点的数据点的条件均值。"自相合"的特性使得主曲线与回归分析完全不同,它没有事先给定的基函数(或假定一个分布),因而能真实反映数据的形态。可以说,主曲线是数据集合的"骨架",数据集合是这个曲线的"云"。

Hastie和Stuetzle(简记为HS)首次提出了主曲线的概念,其定义如下[10]。

定义1 令 $X=\{x_1,x_2,\cdots,x_n\}$ 为 n 个样本组成的 d 维数据集,关于 X 的主曲线是一条光滑的关于单变量 t 的曲线:

$$f(t)=(f_1(t),f_2(t),\cdots,f_d(t))=E\{X|t_f(x)=t\} \quad (1)$$

到稿日期:2010-01-20 返修日期:2010-04-07 本文受国家自然科学基金(60773172)资助。

祁云嵩 博士生,副教授,主要研究方向为模式识别、生物信息学,E-mail:qys@ujs.edu.cn;孙怀江 男,教授,博士生导师,CCF会员,主要研究方向为人工智能、模式识别。

式中，t 为实轴上的区间，$t_f(x)$ 为 $f(t)$ 上和 x 最近的点所对应的 t 的值，即：

$$t_f(x)=\sup\{t:\|x-f(t)\|=\inf_{\tau}\|x-f(\tau)\|\} \quad (2)$$

根据以上定义，计算主曲线的算法（HS 算法[10]）可以通过“期望”距离计算及“投影”计算迭代而成。算法在“期望方差距离”下降到某个阈值时停止迭代。“期望方差距离”定义如下。

定义 2 主曲线的期望方差距离（ESE）为样本数据点与主曲线间最近投影距离的平方和，即：

$$ESE=E\{\|X-f(t_f(X))\|^2\} \quad (3)$$

HS 首次提出主曲线的概念之后，由于其暗示的广泛应用前景，不断有学者从不同角度对其进行改善，提出了不同的主曲线算法[9-12]。由于主曲线与主成分的密切联系，主曲线生成算法通常以第一主成分线为初始值。然而，第一主成分线不能反映数据集的拓扑关系，所以，第一主成分未必是算法初始化的最佳选择。并且，已有的算法中，在迭代计算过程中对主曲线上数据点的优化多采用梯度优化法，不能直接体现主曲线的特点。本文采用非线性回归曲线代替第一主成分线来初始化主曲线，在迭代过程中使用顶点优化法来计算主曲线。

设微阵列中某类样本的样本数为 n，每个样本的基因数为 m。以各样本的基因序号为横坐标，各基因的表达值为纵坐标，每个基因的表达值对应于平面坐标系上的一个点 x，则该类样本的主曲线的维数 $d=m*n$。具体主曲线生成算法如下：

1）初始化

采用非线性回归方法，生成初始的“主曲线”$f^{(0)}(t)$，令 $j=0$。

2）投影

根据主曲线自相合的性质，设第 j 次迭代结果为 $f^{(j)}(t)$，新的投影坐标为：

$t_i^{(j+1)}=t_{f^{(j)}}(x_i)$，$i=1,2,\cdots,d$。投影计算如下：

(a) 记 d_{ik} 为 x_i 到各区间 $[t_k^j,t_{k+1}^j]$（$k=0,1,\cdots,d-1$）上的主曲线的距离，对应的投影点参数为 t_k^*；

(b) $t_i^{(j+1)}$ 取值为最小的 d_{ik} 所对应的 t_k^*；

(c) 计算 $t_i^{(j+1)}$ 对应的 $f^{(j+1)}$ 值；

(d) $t_i^{(j+1)}$ 更新为 $f_1^{(j)}(t)$ 到 $f_i^{(j)}(t)$ 间的弧长。

3）顶点优化

令 X 表示基因数据点的集合，V_i 表示各数据点的近邻集，即：$V_i=\{x\in X:\|x-x_i\|<\|x-f(t)\|,i=1,2,\cdots,n\}$。记各 V_i 所对应的数据中心为 C_{V_i}。假设其他点不动，将主曲线上离 C_{V_i} 最近的 $f_i(t)$ 点移到 C_{V_i}，计算所对应的 ESE 的减少量 ΔESE。最后，将最大 ΔESE 所对应的 $f_i(t)$ 移向对应的 C_{V_i}，移动的距离为 $\xi\cdot\|C_{f(\lambda_k)}-C_{V_i}\|$。其中，实验确定系数 ξ 取值为 0.36 时算法取得较好的性能。

4）迭代

重复步骤 2）和步骤 3），直到 ESE 不再下降或下降幅度小于指定的阈值为止。

3 基于主曲线的微阵列数据分类

将微阵列数据中每一个基因的表达值都看作是一个数据点，计算训练样本中各类样本的主曲线，然后，计算测试样本上各点到各类样本主曲线的 ESE 值，则最小 ESE 值所对应的主曲线所属的类别即为测试样本的测试类别。

4 实验及结果分析

为了测试上述算法对微阵列数据分类的性能，我们选择了几个典型的算法进行比较。这些算法有支撑向量机分类法（SVM）、贝叶斯分类法（NB）以及 K 近邻分类法（k-NN）。其中，SVM 通过在高维空间中寻找一个将类间隔最大化的超平面来对样本进行分类；NB 通过计算样本属于各类别的后验概率来确定样本所属的类别；k-NN 分类法根据样本间的相似度来对样本进行分类。现有文献资料表明，这些算法均能成功地用于微阵列数据的分类。

如表 1 所列，在实验中，我们根据其他研究文章选用了一些典型的微阵列数据集进行试验。这些数据集在样本类别和每类样本数量两个方面均具有一定的代表性。而样本类别数以及各类样本的样本数均能影响分类器的性能。

表 1 几种典型的微阵列数据集

微阵列数据集	样本数	基因数	样本类别数
DLBCL	77	7070	2
Prostate	102	12533	2
SRBCT	83	2308	4
NCI-60	60	5726	9

图 1 展示了 4 个实验数据集上的各类别样本的主曲线图。从主曲线图上可以看出，DLBCL 数据集、Prostate 数据集以及 SRBCT 数据集上的主曲线相互间距较大，能很好地区分开各类样本，而 NCI 数据集中，9 类样本间的主曲线间距较小，部分主曲线甚至交织在一起，直观上体现了样本分类的困难。表 2 所列的分类准确率也说明了这一现象。

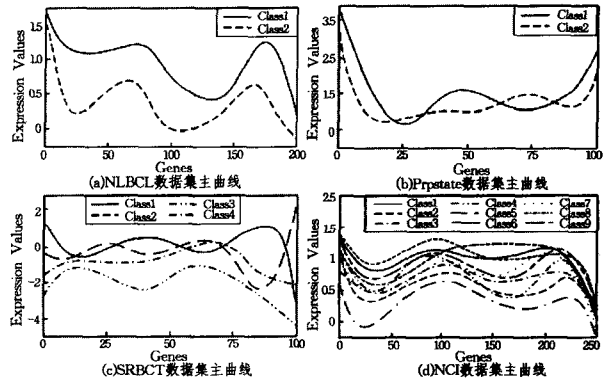

图 1 实验数据集上的主曲线

表 2 PC 分类法与其他主流分类法的分类性能比较

微阵列数据集	分类准确率			
	PC	SVM	k-NN	NB
DLBCL	93.03%	96.56%	88.10%	84.26%
Prostate	94.03%	94.38%	83.97%	85.42%
SRBCT	98.35%	99.12%	87.89%	76.11%
NCI-60	82.51%	81.67%	78.32%	74.39%

表 2 列出了 PC 分类法与其他几种主流分类法在上述 4 个实验数据集上的分类准确率。虽然 PC 分类法在 DLBCL 数据集、Prostate 数据集以及 SRBCT 数据集上的准确率要稍逊于 SVM 分类方法，但在 NCI 数据集上的分类准确率却优

于所有的其他分类方法。应该说，较之其他几种数据集，对NCI数据集的分类更具有挑战性。NCI数据集有9个类别且总共只有60个样本，每类样本的样本数只有2～9个。在这样一个小样本数据集上，基于复杂数学模型的分类方法均不可避免地存在"过拟合"现象。虽然有研究表明SVM分类方法很适合微阵列数据分类处理，但在多类别、超小样本数据集上，本文的PC分类法表现出了一定的优异性。

结束语 本文提出了一种基于主曲线的微阵列数据分类方法，通过计算测试样本与各类样本主曲线的方差距离来确定其类别。实验结果表明，在处理高维小样本数据时，本文的方法较之其他几种主流分类方法有一定的优势。微阵列数据是典型的高维、小样本数据，其高维特性可以通过基因选择方法来进行降维处理，但其小样本特性却是我们在分类器设计时必须面对的问题，所以，本文的分类方法是一个值得继续深入研究的课题。

参 考 文 献

[1] Pochet N, De Smet F, Suykens J A, et al. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction[J]. Bioinformatics, 2004, 20(17): 3185-3195

[2] Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges[J]. ACM SIGKDD Explorations Newsl etter, 2003, 5(2): 1-5

[3] Li J, Ng S, Wong L. Bioinformatics adventures in database research[C]// Proceedings of the international conference on database theory (ICDT). 2002: 31-46

[4] Cho S-B. Exploring features and classifiers to classify gene expression profiles of acute leukemia[J]. Int. J. Pattern Recogn. Artif. Intell, 2002, 16(7): 1-13

[5] Eisen M B, Brown B O. DNA arrays for analysis of gene expression[J]. Methods Enzymol, 1999, 303: 179-205

[6] Mallick B K, Ghosh D, Ghosh M. Bayesian classification of tumors using gene expression data[J]. Journal of the Royal Statistical Society, 2005, B 67: 219-232

[7] Quinlan J R. C4. 5: Programs for Machine Learning[M]. San Mateo, California: Morgan Kaufmann Publishers, 1993

[8] A novel ensemble of classifiers for microarray data classification

[9] Hastie T, Stuetzle W. Principal curves[J]. J. A. Stat. A, 1988, 84: 502-516

[10] Hastie T. Principal curves and surfaces[D]. Standford University, 1984

[11] Kegl B, Krzyzak A, Linder T, et al. Learning and Design of Principal Curves[J]. IEEE Trans. Pattern. Anal. Mach. Intell., 2000, 22(3): 281-297

[12] Wang Haonan, Lee Thomas C M. Extraction of Curvilinear Features from Noisy Point Patterns using Principal Curves[J]. Pattern Recognition Letters, 2008, 29: 2078-2084

(上接第192页)

参 考 文 献

[1] Fonseca C M, Fleming P J. Genetic algorithms for multiobjective optimization: formulation, discussion and generation[A]// Proceedings of the 5th International Conference on Genetic Algorithms[C]. San Mateo, California, 1993: 416-423

[2] Horn J, Nafpliotis N, Goldberg D E. A Niched Pareto Genetic Algorithm for Multiobjective Optimization[A]// IEEE World Congress on Computational Intelligence[C]. Piscataway, New Jersey, 1994: 82-87

[3] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197

[4] Zitzler E, Laumanns M, Thiele L. SPEA2: Improving the Strength Pareto Evolutionary Algorithm[R]. Switzerland, May 2001

[5] Knowles J D, Corne D W. Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy[J]. Evolutionary Computation, 2000, 8(2): 149-172

[6] Van Veldhuizen D A, Lamont G B. Multiobjective Optimization with Messy Genetic Algorithms[A]// Proceedings of the 2000 ACM Symposium on Applied Computing[C]. Villa Olmo, Como, Italy: ACM, 2000: 470-476

[7] Coello Coello C A, Van Veldhuizen D A, Lamont G B. Evolutionary Algorithms For Solving Multi-objective Problems[M]. New York: Kluwer Academic/Plenum, 2002

[8] Coello Coello C A. Evolutionary multi-objective optimization: a historical view of the field[J]. Computational Intelligence Magazine, IEEE, 2006, 1(1): 28-36

[9] 刘旭红，刘玉树，张国英，等. 多目标优化算法NSGA-II的改进[J]. 计算机工程与应用，2005(15): 73-75

[10] Ciekki Coello C A. List of References on Evolutionary Multiobjective Optimization[EB/OL]. http://www.lania.mx/ ~ccoella/EMOObib.html

[11] Branke J, Schmeck H, Deb K, et al. Parallelizing Multi-objective Evolutionary Algorithms: Cone Separation[A]// Congress on Evolutionary Computation(CEC'2004)[C]. IEEE, 2004

[12] Deb K, Zope P, Jain A. Distributed computing of pareto-optimal solutions with evolutionary algorithms[J]. Evolutionary Multi-Criterion Optimization, 2003, 2632: 534-549

[13] Ishibuchi H, Narukawa K. Spatial Implementation of Evolutionary Multiobjective Algorithms with Partial Lamarckian Repair for Multiobjective Knapsack Problems[A]// Proceedings of the Fifth International Conference on Hybrid Intelligent Systems[C]. IEEE, 2005

[14] 周育人，李元香，王勇. 一种有效的实数编码遗传算法[J]. 武汉大学学报：理学版，2003, 49(1): 39-43

[15] Zitzler E, Deb K, Thiele L. Comparison of multiobjective evolutionary algorithms: Empirical results[J]. Evol. Comput, 2000, 8(2): 173-195

[16] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach[J]. IEEE Transactions on Evolutionary Computation, 1999, 3(4): 257-271