癌症基因分类的 Laplace 谱方法

王 年,庄振华,范益政,李学俊,王 继
(安徽大学计算智能与信号处理教育部重点实验室,安徽合肥 230039)

摘 要: 本文尝试着将图的 Laplace 谱理论应用于癌症基因表达谱数据的分类上.计算出训练集中每个类的均值作为类中心,选出与类中心欧式距离最小的若干样本用 laplace 矩阵构造完全图,记为代表该类的标准图.用待测样本依次替换标准图中所有的点,将生成的新图与标准图进行特征点匹配,并计算匹配点数总和.将待测样本划分为总匹配点数最多的那个类.通过对白血病两个亚型(ALL 与 AML)与结肠癌数据进行留一法实验,验证了本文方法的有效性.

关键词: 分类; 基因表达谱数据; Laplace 谱
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2011) 07-1594-04

Classification of Tumor Gene Expression Data Based on Laplacian Spectra of Graphs

WANG Nian, ZHUANG Zhen-hua, FAN Yi-zheng, LI Xue-jun, WANG Ji
(*Education Ministry Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei, Anhui* 230039, *China*)

Abstract: We introduce a novel classification algorithm for gene expression data based on the Laplacian spectra of graphs. The class center is obtained by computing the average of each class in the traning set, and the Laplacian matrices of complete graphs so called normal graphs are constructed on some samples with the minimum Euclidean distance between the class center. The sum of matched points are calculated by replacing points of standard image with test samples. The test sample is divided into the biggest one of the total matched points of the class. The effectiveness of this algorithm has been verified through the leaving-one experiments using Leukemia data and Colon cancer data.

Key words: classification; gene expression data; laplace spectra

1 引言

在癌症的诊疗过程中,根据病人的生理状况以及癌症类型选择正确的医疗方案是治疗成功的前提.然而癌症种类繁多,即便是同种癌症的不同亚型也可能对应不同的治疗方案.因此,如何准确地对癌症进行分类成为癌症治疗中需要迫切解决的问题.DNA 微阵列技术的出现使得人们可以从分子角度对癌症展开更加深入的研究,同时也积累了大量的基因表达谱数据.

由于基因表达谱数据具有样本少、维数高的特点,在数据维数远大于样本个数的情况下运用机器学习的方法进行分类显得十分困难.因此,如何有效地从基因表达谱数据中挑选出蕴含大量分类信息的特征基因显得尤为重要.Golub[1]等人于 1999 年提出了以"信噪比"作为特征提取指标,用投票表决法对白血病的两个亚型进行了分类研究.此后,Singh. D[2]等人选用了与 Golub 相同的特征提取指标,用 K 近邻法作为分类方法对前列腺癌基因进行了分类研究.2002 年,Guyon[3]等人使用支持向量机(SVM)迭代去除对构成分类超平面的元素影响最小的基因,然后同样使用支持向量机(SVM)进行了分类实验.用于研究基因表达谱数据分类的方法还有分层聚类法[4,5],贝叶斯决策[6,7],人工神经网络[8],决策树[9,10],遗传算法[11]以及基于权重的关联空间分类模型[12]等.

传统的机器学习方法首先对基因表达谱数据进行特征基因提取,然后将提取出来的特征基因作为分类特征输入分类器以达到分类目的.然而基因表达谱数据所具有的高维数、低样本的特性,使得传统方法很难从中提取出有用的特征信息,往往无法取得较好的分类效果.本文尝试着将图的 Laplace 谱理论应用到基因表达谱数据的分类上来,将数据集中的每个样本作为图中的特征点,而样本中的基因表达信息则作为该特征点的

收稿日期:2010-03-01;修回日期:2011-01-26
基金项目:国家自然科学基金(No.60772121)

坐标信息,通过分析其不同点集在空间中的结构异同点来进行癌症基因表达谱数据的分类.

本文所提出的分类方法是计算出训练集中每个类的均值作为类中心,选出训练集中与类中心欧式距离最小的若干个样本作为特征点集,构造 Laplace 完全图,并记为代表该类的标准图.然后用待测样本逐个替换标准图中所有的点,将生成的新图与标准图进行特征点匹配,计算匹配点数总和.最后将待测样本划分为总匹配点数最多的那个类.通过对白血病两个亚型(ALL 与 AML)和结肠癌的基因表达谱数据应用本文方法进行了留一法实验,获得了良好的效果,并和传统方法进行了对比,证明了本文方法的有效性.

2　基于图的 Laplace 谱的点匹配算法

对含有 m 个特征点 $p_i(i=1,2,\cdots,m)$ 的图 I,其高斯权 Laplace 矩阵定义为:

$$\boldsymbol{L}=[l_{i,j}]=\begin{cases}\exp\left(-\dfrac{\|p_i-p_j\|^2}{2\sigma^2}\right), & i\neq j\\ -\sum\limits_{k\neq i}l_{i,k}, & i=j\end{cases}\tag{1}$$

其中,参数 σ 可根据点的特征来选取.

对高斯权 Laplace 矩阵 $\boldsymbol{L}$ 进行 SVD 分解,得到

$$\boldsymbol{L}=\boldsymbol{U}\Delta\boldsymbol{U}^{\mathrm{T}}\tag{2}$$

其中 $\Delta=\mathrm{diag}\{\lambda_1,\lambda_2,\cdots,\lambda_i,\cdots,\lambda_m\}$,$\lambda_i(i=1,2,\cdots,m)$ 是 $\boldsymbol{L}$ 的奇异值,$\boldsymbol{U}=(u_1,u_2,\cdots,u_i,\cdots,u_m)$ 是 m 阶正交矩阵,$\boldsymbol{u}_i(i=1,2,\cdots,m)$ 是 $\boldsymbol{U}$ 的列向量.

设待匹配的图 I 与 J 经式(1)、(2)变换,得到 m 阶正交矩阵 $\boldsymbol{U}_I$,$\boldsymbol{U}_J$,记 $\boldsymbol{U}_I$ 与 $\boldsymbol{U}_J$ 的第 i 行为 u_I^i 和 u_J^i,它们分别是图 I 与 J 的第 i 个点的特征表示.构造如下匹配矩阵:

$$\boldsymbol{D}=[d_{i,j}]=\boldsymbol{U}_I\boldsymbol{U}_J^{\mathrm{T}}=[\boldsymbol{u}_I^i(\boldsymbol{u}_J^j)^{\mathrm{T}}]\tag{3}$$

若 $\boldsymbol{u}_I^i$ 与 $\boldsymbol{u}_J^j$ 之间的距离越小,$d_{i,j}=\boldsymbol{u}_I^i(\boldsymbol{u}_J^j)^{\mathrm{T}}$ 的值则越大,这就意味着图 I 的第 i 个特征点与图 J 的第 j 个特征点匹配的可能性越大.因此,如果 $d_{i,j}$ 是它所在的行与列的最大值,则认为图 I 的第 i 个特征点与图 J 的第 j 个特征点匹配[13].

3　基于 Laplace 谱的癌症基因分类算法

基因表达谱数据一般由若干个样本集合而成,可将某个样本记为 $G=\{g_1,g_2,\cdots,g_n\}$,其中 g_i 代表该样本的第 i 个基因在该次实验中的表达等级.由此可见单个样本的基因表达谱数据是一个数字序列,并不能表达结构信息,如果用图谱的方法来解决基因表达谱数据分类的问题,就可以将同一类样本构造为一个图,而该图的结构信息反映了本类样本的特征.

设基因表达谱数据集中包含类别 $A=\{a_1,a_2,\cdots,a_i,\cdots,a_q\}$ 与类别 $B=\{b_1,b_2,\cdots,b_i,\cdots,b_p\}$ 两个集合,其中 $a_i(i=1,2,\cdots,q)$ 是 A 中的样本,$b_i(i=1,2,\cdots,p)$ 是 B 中的样本,每个样本包含 n 个基因表达等级数据.基因表达谱数据集中的每个样本代表 n 维空间中的一个点,不同类的点集间具有不同的空间分布,或者不同的空间结构,因此不同点集所构成的图包含了大量的分类信息.

本文算法如下:

Step 1　从数据集中随机挑选出一个样本作为待测样本 t,剩余样本作为训练样本.训练样本数据集分为两类,记为 A 类与 B 类.

Step 2　计算类中样本均值 $mean_A=\dfrac{\sum\limits_{i=1}^{q}a_i}{q}$ 与 $mean_B=\dfrac{\sum\limits_{i=1}^{p}b_i}{p}$,其中 $a_i(i=1,2,\cdots,p)$ 为 A 类中样本的基因表达等级,$b_i(i=1,2,\cdots,q)$ 为 B 类中样本的基因表达等级.

Step 3　从 A 类中挑选出 m 个基因表达等级与 $mean_A$ 欧式距离最小的样本,称这 m 个样本为 A 类的特征点,利用这 m 个样本构造一个完全图,记为标准图 I.用同样的方法处理 B 类,挑选出 m 个基因表达等级与 $mean_B$ 欧式距离最小的样本作为 B 类的特征点,可以得到标准图 J.

Step 4　用待测样本 t 依次替换标准图 I 与 J 中的第 $i(i=1,2,\cdots,m)$ 个样本(即图的第 i 个节点)得到两个图序列,记为 $I_i(i=1,2,\cdots,m)$ 与 $J_i(i=1,2,\cdots,m)$.

Step 5　利用 Laplace 谱的点匹配算法,将 $I_i(i=1,2,\cdots,m)$ 依次与标准图 I 进行匹配,记每次正确匹配的点数为 $x_i(i=1,2,\cdots,m)$.用同样的方法将图 $J_i(i=1,2,\cdots,m)$ 依次与标准图 J 进行匹配,得到 $y_i(i=1,2,\cdots,m)$.

Step 6　计算 $X=\dfrac{\sum\limits_{i=1}^{m}x_i}{m}$ 与 $Y=\dfrac{\sum\limits_{i=1}^{m}y_i}{m}$,若 $X>Y$,则待测样本为 A 类;若 $X<Y$,则待测样本为 B 类;若 $X=Y$,则考虑样本错判代价:在正常样本与患病样本的二分类实验中,当 $X=Y$ 时,待测样本被划分为癌症样本,因为将癌症样本错判为正常样本的代价远大于将正常样本错判为癌症样本.在多种癌症数据参与分类的实验中,当 $X=Y$ 时,待测样本被划分为拒绝类,即不属于训练集中的任何一类.

Step 7　重复 step1 至 step 6 的步骤,直到数据集中的每个样本都被作为待测样本使用过一次为止,然后统计该方法在留一法实验中的分类正确率.

4 实验结果与分析

本文使用白血病两个亚型(ALL 和 AML)以及结肠癌的基因表达谱数据进行了留一法实验.白血病数据包括 52 个样本,每个样本包含 12564 个基因表达谱数据,其中 24 个样本为 ALL,另外 28 个样本为 AML(数据来自于 http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi).结肠癌数据包括 62 个样本,每个样本包含 2000 个基因表达数据,其中正常样本为 22 个,患病样本为 40 个(数据来自于 http://linus.nci.nih.gov/~brb/DataArchive_New.html).留一法实验即每次从数据集中选取一个样本作为待测样本,其余样本作为训练样本进行分类实验,直到数据集中每个样本都被作为待测样本使用过一次为止.这种方法的训练集和测试集相互独立,同时又能避免单次留取实验所造成的偏性,保证了实验的客观性.

本文将从分类准确率、运算复杂度及鲁棒性三个方面对算法进行分析.这是因为:

(1)基因表达谱具有低样本特性,因此如何在低样本的条件下获得较好的分类效果将是基因表达谱分类算法必须解决的最主要问题.

(2)基因表达谱具有高维数特性,使用传统的分类方法将面对维数灾难问题,良好的实时性将是基因表达谱分类算法运用于临床实践的重要保证.

(3)基因表达谱数据由 DNA 微阵列实验获得,实验误差及操作失误将使得数据中包含大量的随机噪声和异常值,这将是对算法鲁棒性的巨大考验.

为评估本文算法在分类准确率、时间复杂度及鲁棒性方面的表现,我们还选用了另外两种基因表达谱数据分类算法与之相比较.这两种算法使用相同的特征提取指标,即 Golub[1]等人所提出的"信噪比"(S2N)指标,分别以 K 近邻法(KNN)[2]和支持向量机(SVM)作为各自的分类器,下文中记为 S2N_KNN 法和 S2N_SVM 法.

本文的实验结果是在配置酷睿双核 1.8G CPU 及 2G 内存的计算机上得到的.由表 1 中可以看出,本文的方法在两个数据集中较另外两种算法都能得到较好的分类正确率及较快的运算时间.这是由于所采用的基于 Laplace 谱的特征点匹配算法能准确检测出加入待测样本后的图结构变化,且该算法只需将各维数据进行简单的加权运算,其运算速度快,运算复杂度不会随维数增加而以几何级数递增,避免产生维数灾难.

图 1 和图 2 分别是白血病数据和结肠癌数据分类准确率随信噪比的变化曲线图,可见,在加入高斯白噪声的情况下,本文的方法相对于传统的分类方法有着更好的正确率与稳健性.这是因为本文算法能够自动剔除训练集当中离类中心较远的样本,降低了数据集中异常值及实验误差对实验结果所造成的影响,具有更强的鲁棒性.传统算法过分依赖于样本的数量,研究者们希望将更多的样本信息输入分类器,以获得更好的分类结果.而本文算法对样本数量的依赖不强,在剔除不良样本之后,具有更好的正确率与稳健性.

表 1 实验结果对比

数据集	实验方法	分类正确率	运算时间
ALL 与 AML	本文方法	98.08%	5.133s
ALL 与 AML	S2N_KNN	76.92%	68.281s
ALL 与 AML	S2N_SVM	89.47%	7.132s
结肠癌	本文方法	88.71%	0.955s
结肠癌	S2N_KNN	85.48%	33.3990s
结肠癌	S2N_SVM	79.03%	10.043s

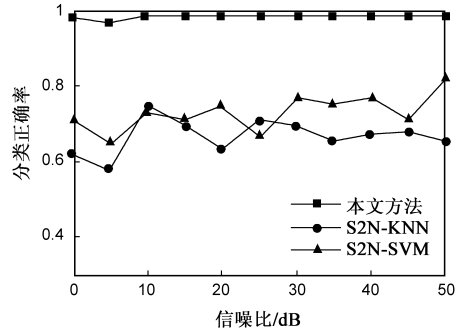

图1 白血病数据中分类正确率随信噪比变化的曲线

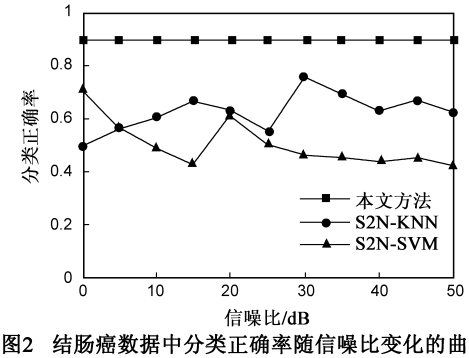

图2 结肠癌数据中分类正确率随信噪比变化的曲线

图 3 与图 4 表明选取合适的特征点个数可以得到较高的分类正确率.在谱图方法中,特征点的选取数量与匹配结果是息息相关的,这是因为如果所取的特征点数太少,分类信息就不足,算法无法较为完整地提取

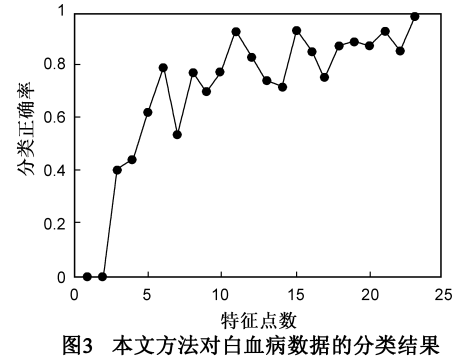

图3 本文方法对白血病数据的分类结果

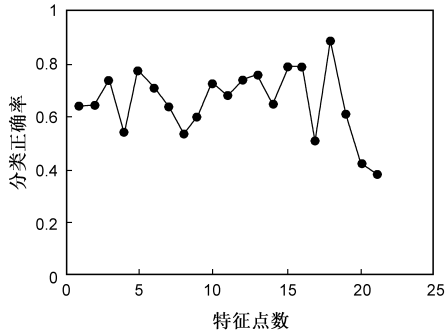

图4　本文方法对结肠癌数据的分类结果

出该类的特征;当所取的特征点数太多时,又会包含一些噪声较大的点,这些点不仅会降低分类正确率,还会徒增算法运算量.因此如何选取合适的特征点数量,加强算法的自适应能力将是以后的重点研究工作.

5　结论

利用基因表达谱数据进行癌症的分类与识别是当前生物信息学研究的主要方向之一.本文尝试着将基于图的 Laplace 谱理论应用于癌症基因分类,在针对白血病的两个亚型(ALL 与 AML)与结肠癌的基因表达谱数据分类实验中都取得了较好的效果.实验结果证明,用该方法进行基因表达数据的分类有着较高的分类正确率及稳健性.由于基于图谱理论的分类过程较为复杂繁琐,并且对特征点的要求也比较苛刻,因此本文的进一步工作就是设法提高算法的运算效率及自适应能力.

参考文献

[1] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531 – 537.

[2] Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002, 1(2): 203 – 209.

[3] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. MachineLearning, 2002, 46(1 – 3): 389 – 422.

[4] Michael B Eisen, Paul T Spellman, Patrick O Borwn, et al. Cluster analysis and dispay of genome wide expression pattenrs [J]. PNAS USA, 1998, 95(25): 14863 – 14868.

[5] Brazma A, Vilo J. Gene expression data analysis[J]. FEBS Letters, 2000, 480(1): 17 – 24.

[6] Anderw D Keller, Michel S chummer, Lee Hood, et al. Bayesian Classification of DNA Array Expression Data[R]. Technical Report UW-CSE-2000-08-01, Department of Computer Science & Engineering, University of Washington, Seattle, 2000.

[7] Zhou Xiaobo, Wang Xiaodong, Dougherty ER. A Bayesian approach to nonlinear porbit gene selection and classification[J]. Journal of the Franklin Institute, 2004, 341(1 – 2): 137 – 156.

[8] Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature Medicine, 2001, 7: 673 – 679.

[9] Zhang Heping, Yu Chang-Yung, Singer Burton, et al. Recursive partitioning for tumor classification with gene expression micoarray data[J]. PNAS USA, 2001, 98(12): 6730 – 6735.

[10] 李颖新,刘全金,阮晓钢.一种肿瘤基因表达数据的知识提取方法[J].电子学报,2004,32(9):1479 – 1482.
LI Ying-xin, LIU Quan-jin, RUAN Xiao-gang. A method for extracting knowledge from tumor gene expression data[J]. Acta Electronica Sinica, 2004, 32(9): 1479 – 1482. (in Chinese)

[11] 蔡立军,林亚平,卢新国,易叶青,李小龙.基于遗传算法的基因分类[J].电子学报,2006,34(11):2115 – 2119.
CAI Li-jun, LIN Ya-ping, LU Xin-guo, YI Ye-qing, LI Xiao-long. Gene clustering based on genetic algorithm[J]. Acta Electronica Sinica, 2006, 34(11): 2115 – 2119. (in Chinese)

[12] 卢新国,林亚平,王海军,李小龙,易叶青.基于微阵列基因表达谱的一种关联空间的癌症分类算法[J].电子学报,2008,36(4):614 – 619.
LU Xin-guo, LIN Ya-ping, WANG Hai-jun1, LI Xiao-long, YI Ye-qing. A relative space based cancer classification with gene expression profiles[J]. Acta Electronica Sinica, 2008, 36(4): 614 – 619. (in Chinese)

[13] 王年,范益政,韦穗,梁栋.基于图的 Laplace 谱的特征匹配[J].中国图象图形学报,2006,11(3):332 – 336.
WANG Nian, FAN Yi-zheng, WEI Sui, LIANG Dong. Feature Matching Based on Laplace an spectra of graphs[J]. Journal of Image and Graphics, 2006, 11(3): 332 – 336. (in Chinese)

作者简介

王　年　男,1966 年生,安徽省和县人,1986 年毕业于安徽大学电子工程与信息科学系,2005 年在该校获博士学位,现为安徽大学教授,主要研究领域为计算机视觉、图像处理、模式识别与生物信息学等,已发表学术论文 50 多篇.
E-mail: wn_xlb@ahu.edu.cn

庄振华　男,1984 年生于福建漳州,2007 年获华侨大学工学学士学位,现为安徽大学电子科学与技术学院硕士研究生,主要研究方向为模式识别与生物信息学.
E-mail: zhzhuang1016@yahoo.com.cn