基于微阵列基因表达谱的一种关联空间的癌症分类算法

卢新国[1], 林亚平[2,1], 王海军[1], 李小龙[1], 易叶青[1]
(1. 湖南大学计算机与通信学院, 湖南长沙 410082; 2. 湖南大学软件学院, 湖南长沙 410082)

摘 要: 利用微阵列基因表达谱分类癌症患者样本对患者的治疗具有非常重要的意义. 针对高维、高冗余的微阵列基因数据中致癌因子存在局部相关性的特点, 提出一种基于权重的关联空间分类模型(Weight based Classification with Related Space, WCRS). 基本思想是首先利用协方差矩阵的对角化来构建癌症组的关联空间, 并提出一种基于癌症关联空间的基因表达模式, 然后提取使得癌症组具有最小组能量的最小扩展空间, 最后在最小扩展空间上建立一种基于权重的癌症分类算法. 实验结果表明, WCRS 在精确度上比传统分类算法具有更好的性能.

关键词: 癌症分类; 基因表达谱; 关联空间
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2008) 04-614-06

A Relative Space Based Cancer Classification with Gene Expression Profiles

LU Xin-guo[1], LIN Ya-ping[2,1], WANG Hai-jun[1], LI Xiao-long[1], YI Ye-qing[1]
(1. *School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China;*
2. *School of Software, Hunan University, Changsha, Hunan 410082, China*)

Abstract: Classification of patient samples with gene expression profiles is important to cancer treatment. In the large redundant and high dimensional gene expression data, a cancer is sensitive to some cancerogenic factors while another cancer is sensitive to some others. So we proposed a weight based classification with relative space(WCRS). The main idea is that a cancer's relative space is obtained via the diagonalization of its covariance matrix, and we built the cancer's model based on its relative space. Then the energy of a cancer is presented for measuring its relative spaces, and a minimal spread space based classification algorithm is proposed. The experiments show WCRS makes better precision than traditional classifications.

Key words: cancer classification; gene expression profile; relative space

1 引言

对癌症的精确分类是提高癌症诊断准确率和治愈率至关重要的一个环节. 近年来, 研究人员利用基因微阵列(Microarray)技术获得大规模基因表达谱数据(Gene Expression Profiles). 一方面, 在高维基因表达谱中存在大量噪声; 另一方面, 在大规模的基因表达数据中存在大量在分类学意义上的冗余基因. 如何利用这种具有高维、高噪、高相关(冗余)特点且只包含有限样本的基因表达谱数据, 识别对疾病有鉴别意义的特征基因或与疾病相关的基因, 为机器学习研究提出了新的课题[1,2].

降维方法经常被利用来处理高维、高噪问题. 对于癌症检测与分类, 一类主要的降维方法是特征选择法, 包括信噪比(Signal to Noise Ratio, SNR)[3], 邻居分析(Neighborhood Analysis)[4], 排秩法(Rank)[5]等. 由于不同的特征选择法使用的搜索机制和评价策略不同, 挑选出的特征基因明显不同[6]. 并且在高冗余的基因表达谱数据中选取数量有限的基因容易丢失有生物价值和分类意义的信息. 另一类降维方法是特征变换法. Conde 等利用聚类方法[7,8]产生基因簇, 然后用簇的平均值训练检测癌症的感知器模型[9]. Khan 等利用主分量分析法(Principal Components Analysis, PCA)研究儿童小圆形蓝色细胞恶性肿瘤 4 种亚型的癌症识别[10]. 建立在降维方法之上的癌症分类模型的基本思想是通过特征选取或特征变换法获取基因特征, 然后利用获取的特征来训练分类器以识别测试样本的癌症类型. 但是目前的癌症分类模型都是利用统一的基因特征建立分类器, 没有充分考虑在生物意义上致癌因子存在局部相关性, 即某一

收稿日期: 2006-08-28; 修回日期: 2007-09-19
基金项目: 湖南省自然科学基金(No. 07JJ5085)

癌症组与一些致癌因子的表达非常相关,而另一癌症组则与另外一些致癌因子的表达非常相关.

本文通过特征变换法来揭示每一组癌症表达数据中隐含变量的空间结构,选择具有最小扩展性的隐含变量组成这一组癌症的最小扩展空间,一组癌症的最小扩展空间具有使得该组癌症样本组能量最小的特性,并提出一种基于权重的关联空间分类模型.在基于权重的关联空间分类模型中每组癌症都具有各自相关的基因特征,不同癌症组的分类器建立在不同的基因特征子集之上.每组癌症的最小扩展空间揭示出隐含的致癌因子和基因在癌症组中的表达,从癌症生物病理上有效地排除噪声和冗余的干扰.

2　问题描述

基因表达谱数据是典型的高维数据,在微阵列数据中包含有数目巨大的变量(成千上万的基因),但只有有限的几十个样本.影响利用基因表达谱进行癌症分类的主要问题是"维数灾难"(Curse of Dimensionality)问题,即有限的对象被分散到高维空间,造成对象(样本)的分布非常稀疏,基因特征的扰动对识别样本类型没有任何影响[11].

影响癌症分类的另外一个因素是在基因数据集中,致癌因子存在局部相关性,即不同癌症组具有不同的致癌因子.如图1所示,数据集合中有两个癌症模式P、Q,癌症P中样本在 $x-y$ 平面图中相互邻近,癌症Q中样本在 $x-z$ 平面图中相互邻近,但在 $x-y$ 平面图中并不能发现癌症Q,在 $x-z$ 平面图中也不能发现癌症P.同时,在所有维空间中也不能发现这两个模式.目前的癌症分类模型都是利用统一的基因特征建立分类器,没有充分考虑生物学意义上致癌因子存在的局部相关性.

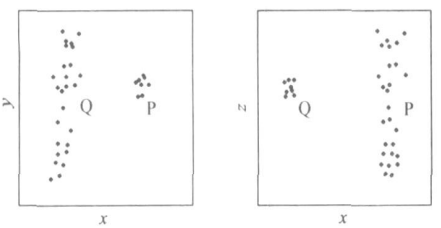

图1　癌症模式P和Q中致癌因子的局部相关性

3　基于关联空间的癌症分类算法

微阵列基因表达谱数据是典型的高维、高噪和高相关数据.传统的癌症分类方法利用降维方法来排除"维数灾难"的影响,但是没有从癌症病理上考虑致癌因子存在的局部相关性.本节构造了癌症的关联空间,以揭示基因特征与癌症之间的相关性以及基因在癌症样本中的表达,并提出了一种基于权重的关联空间癌症分类算法.

3.1　预备知识

设 $\tilde{X}$ 是一组对象集合, $\tilde{X}=\{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n\}$, $\bar{\boldsymbol{x}}_i=(x_{i1}, x_{i2}, \cdots, x_{im})^{\mathrm{T}}$, $\boldsymbol{X}$ 是由 $\tilde{X}$ 中对象组成的矩阵,

$$\boldsymbol{X}=(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \cdots, \bar{\boldsymbol{x}}_n)=\begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix}$$

定义**1**(质心)　对于集合 $\tilde{X}$, $M(\tilde{X})$ 表示的 $\tilde{X}$ 质心, $M(\tilde{X})=1/n\sum_{i=1}^{n}\bar{\boldsymbol{x}}_i$.

定义**2**(迹)　如果 $\boldsymbol{X}$ 是方阵,即 $m=n$ 时, $\mathrm{TR}(\boldsymbol{X})$ 表示 $\boldsymbol{X}$ 的迹, $\mathrm{TR}(\boldsymbol{X})=\sum_{i=1}x_{ii}$.

3.2　基因微阵列表达谱

DNA微阵列是在一定尺寸的基片(如硅片、玻璃、塑料等)表面固定一系列可寻址的识别分子的点阵,点阵中每一个点都可以视为一个传感器的探头.主要是通过DNA-DNA杂交反应或是蛋白质之间的特异性结合,同时观测数千甚至数万个基因.基因表达谱是指利用DNA微阵列所测定的组织样本中基因的表达水平值,通常利用矩阵形式表示.假设 $\boldsymbol{X}$ 为一个 $m\times n$(通常 $m\gg n$)的基因表达矩阵,矩阵 $\boldsymbol{X}$ 的第 i 行是第 i 个基因在所有观测样本中的表达值,第 j 列是第 j 个样本中所有观测基因的表达值.矩阵 $\boldsymbol{X}$ 的元素 x_{ij} 表示第 i 个基因在第 j 个观测样本中的表达水平.

3.3　提取癌症的关联空间

假设基因微阵列表达谱 n 中个样本属于 k 种不同的癌症,第 i 类癌症样本集合是 $\tilde{C}_i$, $\tilde{C}_i$ 的样本数目为 n_i.根据癌症样本的类别,则可以将矩阵基因表达矩阵 $\boldsymbol{X}$ 进行如下变形:

$$C(\boldsymbol{X})=(\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_k) \qquad (1)$$

其中 $\boldsymbol{C}_i$ 是 $\tilde{C}_i$ 中样本的基因表达矩阵,则 $\boldsymbol{C}_i$ 为 m 行 n_i 列的矩阵.

通过以下的方法来获取 $\tilde{C}_i$ 的关联空间 ε,然后 $\tilde{C}_i$ 中样本在 ε 的方向上映射以降低样本维数. $\boldsymbol{C}_i^{\mathrm{T}}=(\bar{\boldsymbol{g}}_1, \bar{\boldsymbol{g}}_2, \cdots, \bar{\boldsymbol{g}}_m)$,其中 $\bar{\boldsymbol{g}}_j$ 是第 j 个基因在 $\tilde{C}_i$ 样本中的基因表达向量,那么 $\boldsymbol{C}_i^{\mathrm{T}}$ 的协方差矩阵为:

$$Cov(\boldsymbol{C}_i^{\mathrm{T}})=\begin{pmatrix} Var(\bar{\boldsymbol{g}}_1) & Cov(\bar{\boldsymbol{g}}_1, \bar{\boldsymbol{g}}_2) & \cdots & Cov(\bar{\boldsymbol{g}}_1, \bar{\boldsymbol{g}}_m) \\ Cov(\bar{\boldsymbol{g}}_1, \bar{\boldsymbol{g}}_2) & Var(\bar{\boldsymbol{g}}_2) & \cdots & Cov(\bar{\boldsymbol{g}}_2, \bar{\boldsymbol{g}}_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\bar{\boldsymbol{g}}_m, \bar{\boldsymbol{g}}_1) & Cov(\bar{\boldsymbol{g}}_m, \bar{\boldsymbol{g}}_2) & \cdots & Var(\bar{\boldsymbol{g}}_m) \end{pmatrix}$$

其中 $Var(\bar{\boldsymbol{g}}_i)$ 为 $\bar{\boldsymbol{g}}_i$ 的方差, $Cov(\bar{\boldsymbol{g}}_i, \bar{\boldsymbol{g}}_j)$ 为 $\bar{\boldsymbol{g}}_i$, $\bar{\boldsymbol{g}}_j$ 之间的协方差.

$Cov(\boldsymbol{C}_i^{\mathrm{T}})$ 是半正定的 m 维方阵,可以进行如下矩阵分解:

$$Cov(\boldsymbol{C}_i^{\mathrm{T}}) = \sum \lambda_r \bar{\boldsymbol{p}}_r \bar{\boldsymbol{p}}_r^{\mathrm{T}} = \boldsymbol{P}\Lambda\boldsymbol{P}^{\mathrm{T}} \quad (2)$$

其中, λ_r 是 $Cov(\boldsymbol{C}_i^{\mathrm{T}})$ 特征值, Λ 是非负的对角矩阵, 对角线上元素由 $\lambda_r(1 \leqslant r \leqslant m)$ 组成, $\bar{\boldsymbol{p}}_r$ 是 λ_r 对应的特征向量, $\boldsymbol{P} = (\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{p}}_2, \cdots, \bar{\boldsymbol{p}}_m)$.

定义 **3**(关联空间) 设癌症 $\tilde{C}_i$ 和表达矩阵 $\boldsymbol{C}_i$, λ_1, $\lambda_2, \cdots, \lambda_m$, 是 $Cov(\boldsymbol{C}_i^{\mathrm{T}})$ 的特征值, $\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{p}}_2, \cdots, \bar{\boldsymbol{p}}_m$ 是 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 对应的特征向量. 对于 $d \leqslant m$, 则由 $\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{p}}_2, \cdots, \bar{\boldsymbol{p}}_d$ 组成了 $\tilde{C}_i$ 上秩为 d 的关联空间 ε, $\varepsilon = \{\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{p}}_2, \cdots, \bar{\boldsymbol{p}}_d\}$, $\bar{\boldsymbol{p}}_i$ 为 ε 的第 i 维方向, λ_i 称为方向 $\bar{\boldsymbol{p}}_i$ 的方向扩展系数, $\boldsymbol{P} = (\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{p}}_2, \cdots, \bar{\boldsymbol{p}}_d)$ 为 $\tilde{C}_i$ 的关联空间矩阵.

定义 **4**(最小扩展空间) 对于癌症 $\tilde{C}_i$ 和表达矩阵 $\boldsymbol{C}_i$, 假设 $\hat{\varepsilon}$ 是 $\tilde{C}_i$ 的 d 维关联空间, $\lambda_1, \lambda_2, \cdots, \lambda_d$ 是方向扩展系数, 满足 $\max(\lambda_1, \lambda_2, \cdots, \lambda_d) \leqslant \min(\lambda_d, \lambda_{d+1}, \cdots, \lambda_m)$ 时, 称 $\hat{\varepsilon}$ 为 d 维最小扩展空间.

定义 **5**(映射) 对于癌症 $\tilde{C}_i$ 和 d 维关联空间 ε, 设 $\tilde{C}_i = \{\bar{\boldsymbol{s}}_1, \bar{\boldsymbol{s}}_2, \cdots, \bar{\boldsymbol{s}}_{n_i}\}$, $\bar{\boldsymbol{s}}_l = (s_{l1}, s_{l2}, \cdots, s_{lm})^{\mathrm{T}}$, 则 $P(\bar{\boldsymbol{s}}_l, \varepsilon)$ 表示 $\bar{\boldsymbol{s}}_l$ 在 ε 的映射, $P(\bar{\boldsymbol{s}}_l, \varepsilon) = (\bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_1, \bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_2, \cdots, \bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_d)^{\mathrm{T}}$, 其中, $\bar{\boldsymbol{p}}_j$ 为 ε 的第 j 维方向, $\bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_j = \sum_{k=1}^{m} s_{lk} p_{jk}$.

定理 **1** 对于癌症 $\tilde{C}_i$ 和 d 维关联空间 ε, $\tilde{C}_i$ 中癌症样本在 $\bar{\boldsymbol{p}}_j$ ($\bar{\boldsymbol{p}}_j \in \varepsilon$) 上映射的方差等于方向 $\bar{\boldsymbol{p}}_j$ 的扩展系数, 在不同方向上的映射之间不相关.

证明 设癌症 $\tilde{C}_i$ 和表达矩阵 $\boldsymbol{C}_i(\bar{\boldsymbol{s}}_1, \bar{\boldsymbol{s}}_2, \cdots, \bar{\boldsymbol{s}}_{n_i})$. 癌症样本 $\bar{\boldsymbol{s}}_l(1 \leqslant l \leqslant n_i)$ 在 ε 上的映射为 $P(\bar{\boldsymbol{s}}_l, \varepsilon)$, 设 $P(\bar{\boldsymbol{s}}_l, \varepsilon) = (p_1, p_2, \cdots, p_d)^{\mathrm{T}}$, 其中 $p_j = \bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_j = \bar{\boldsymbol{p}}_j^{\mathrm{T}} \bar{\boldsymbol{s}}_l$, $1 \leqslant j \leqslant d$. 不妨假设 $P(\bar{\boldsymbol{s}}_l, \varepsilon)(1 \leqslant l \leqslant n_i)$ 为 d 维正态随机列向量, 则 $Var(p_j) = \bar{\boldsymbol{p}}_j^{\mathrm{T}} \lambda_j \bar{\boldsymbol{p}}_j = \lambda_j$. $Cov(p_j, p_k) = Cov(\bar{\boldsymbol{p}}_j^{\mathrm{T}} \bar{\boldsymbol{s}}_i, \bar{\boldsymbol{p}}_k^{\mathrm{T}} \bar{\boldsymbol{s}}_i) = \bar{\boldsymbol{p}}_j^{\mathrm{T}} Cov(\boldsymbol{C}_i^{\mathrm{T}}) \bar{\boldsymbol{p}}_k = \lambda_k \bar{\boldsymbol{p}}_j^{\mathrm{T}} \bar{\boldsymbol{p}}_k$, 由于 $\bar{\boldsymbol{p}}_j$ 与 $\bar{\boldsymbol{p}}_k$ 正交, 所以 $Cov(p_j, p_k) = 0$, 那么 $P(\bar{\boldsymbol{s}}_i, \varepsilon)(1 \leqslant l \leqslant n_i)$ 在 $\bar{\boldsymbol{p}}_j$ 与 $\bar{\boldsymbol{p}}_k$ 方向上不相关. 证毕.

由定理 1 可知, 癌症样本在 ε 上映射后, 不仅可以降低样本维数, 同时产生的基因特征之间互不相关, 可以消除微阵列表达谱数据的冗余和噪声. 方向扩展系数描述了样本在关联空间 ε 上不同方向上的相异度, 即方向扩展系数越小, 癌症样本在该方向上映射后的相异程度越小, 相似程度越大. 通过定义 4, 样本在最小扩展空间的方向上具有比其它相同秩的关联空间方向上更大的相似度.

3.4 基于关联空间的基因表达

假设在癌症 $\tilde{C}_i$ 中存在 d 个控制因子, 这些控制因子隐含于基因表达谱 C_i 中, 并调控所有基因在 $\tilde{C}_i$ 中的表达. 对于 $\bar{\boldsymbol{s}}_l \in \tilde{C}_i$, $\bar{\boldsymbol{s}}_l$ 在 ε 的映射可用矩阵形式表示为:

$$P(\bar{\boldsymbol{s}}_l, \varepsilon) = \boldsymbol{P}^{\mathrm{T}} \bar{\boldsymbol{s}}_l$$
$$\bar{\boldsymbol{s}}_l = \boldsymbol{P} \cdot P(\bar{\boldsymbol{s}}_l, \varepsilon) \quad (3)$$

其中, $\boldsymbol{P}$ 为 $\tilde{C}_i$ 的关联空间矩阵, $\boldsymbol{P}^{\mathrm{T}}$ 为 $\boldsymbol{P}$ 的转置矩阵. 我们可以这样理解癌症 $\tilde{C}_i$ 的基因表达: 一方面, 关联空间矩阵 $\boldsymbol{P}$ 的行和列分别对应基因和控制因子, 其元素 p_{jk} 表示该癌症中第 j 个基因中第 k 个控制因子的控制量. 另一方面, $P(\bar{\boldsymbol{s}}_l, \varepsilon)$ 的行和列分别对应控制因子和样本, 第 k 个元素表示第 k 个控制因子在该样本中的调控度. 因此, 在样本 $\bar{\boldsymbol{s}}_l$ 中第 j 个基因的表达水平即 $\boldsymbol{C}_i$ 的元素 c_{jl} 为所有遗传控制因子 p_{jk} 和其对应的在该样本中调控度乘积之和.

3.5 选取关联空间的维数

定义 **6**(关联空间距离) 设癌症样本 $\bar{\boldsymbol{s}}_i$, $\bar{\boldsymbol{s}}_j$ 和关联空间 ε, $D(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{s}}_j, \varepsilon)$ 表示样本 $\bar{\boldsymbol{s}}_i$ 和 $\bar{\boldsymbol{s}}_j$ 在 d 维关联空间 ε 上的距离, $D(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{s}}_j, \varepsilon) = \| P(\bar{\boldsymbol{s}}_i, \varepsilon) - P(\bar{\boldsymbol{s}}_j, \varepsilon) \|$, 其中 $\bar{\boldsymbol{p}}_k$ 是 ε 的第 k 维方向.

定义 **7**(组能量) 对于癌症 $\tilde{C}_i$ 和质心 $M(\tilde{C}_i)$, n_i 是 $\tilde{C}_i$ 中的样本数目, $E(\tilde{C}_i, \varepsilon)$ 为 $\tilde{C}_i$ 的组能量, $E(\tilde{C}_i, \varepsilon) = \frac{1}{n_i} \sum_{\bar{\boldsymbol{s}}_l \in \tilde{C}_i}^{n_i} \sum_{l=1} \{D(\bar{\boldsymbol{s}}_l, M(\tilde{C}_i), \varepsilon)\}^2$.

组能量 $E(\tilde{C}_i, \varepsilon)$ 反映的是 $\tilde{C}_i$ 中的癌症样本和 $\tilde{C}_i$ 的质心在 ε 上映射后的距离, 它与癌症样本总体 $\tilde{C}$ 在 ε 上的组能量 $E(\tilde{C}, \varepsilon)$ 有如下性质:

定理 **2** 设癌症样本集合 $\tilde{C}_i$ 和癌症样本总体 $\tilde{C}$, 假设 ε 是 $\tilde{C}_i$ 的关联空间, d 是 ε 的秩, 则 $\lim_{d \to \infty} \frac{E(\tilde{C}_i, \varepsilon)}{E(\tilde{C}, \varepsilon)} = 1$.

证明 设癌症样本总体 $\tilde{C}$ 包含 k 种癌症类型, 癌症 $\tilde{C}_i(1 \leqslant i \leqslant k)$ 的样本数是 n_i, $\tilde{C}$ 的样本数是 n, $n = \sum_{i=1}^{k} n_i$.

由于, $D(\bar{\boldsymbol{s}}_l, M(\tilde{C}_i), \varepsilon) = \sum_{r=1}^{d} \left(\bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_r - M(\tilde{C}_i) \cdot \bar{\boldsymbol{p}}_r\right)^2$,

那么, $E(\tilde{C}_i, \varepsilon) = \frac{1}{n_i} \sum_{l=1}^{n_i} \sum_{r=1}^{d} \left(\bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_r - M(\tilde{C}_i) \cdot \bar{\boldsymbol{p}}_r\right)^2$

$$= \frac{1}{n_i} \sum_{r=1}^{d} \sum_{l=1}^{n_i} \left(\bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_r - M(\tilde{C}_i) \cdot \bar{\boldsymbol{p}}_r\right)^2.$$

同理, $E(\tilde{C}, \varepsilon) = \frac{1}{n} \sum_{i=1}^{k} \sum_{r=1}^{d} \sum_{l=1}^{n_i} \left(\bar{\boldsymbol{s}}_l \cdot \bar{\boldsymbol{p}}_r - M(\tilde{C}_i) \cdot \bar{\boldsymbol{p}}_r\right)^2$.

显然, $\lim_{d \to \infty} \frac{E(\tilde{C}_i, \varepsilon)}{E(\tilde{C}, \varepsilon)} = 1$. 证毕.

定义 **8**(组能量均值) 设癌症 $\tilde{C}$ 包含 k 种癌症类型, $\tilde{C}_i(1 \leqslant i \leqslant k)$ 的关联空间为 ε_i, 则用 $averE(\tilde{C})$ 表示 $\tilde{C}$ 的组能量均值, $averE(\tilde{C}) = \frac{1}{k} \sum_{i=1}^{k} \frac{E(\tilde{C}_i, \varepsilon_i)}{E(\tilde{C}, \varepsilon_i)}$.

如何确定关联空间的维数是选取癌症关联空间非常重要的一个方面. 从定理 2 可以看出, 如果选取的关联空间的维数过高, 则 $\tilde{C}$ 中的癌症样本和 $\tilde{C}_i$ 的质心在 ε_i 映射后的距离接近于癌症样本总体 $\tilde{C}$ 中的癌症样本和 $\tilde{C}$ 的质心在 ε_i 映射后的距离. 同样会导致"维数灾

难". 显然, 当在 ε_i 下可以识别癌症 $\tilde{C}_i$ 时, $averE(\tilde{C}) \ll 1$. 为了避免利用所有不同秩的关联空间来训练分类模型所消耗的大量时间开销, 我们首先确定一个参数阀值 λ $(0< \lambda< 1)$, 然后计算所有维 $aveE(\tilde{C})$, 选择满足 $averE(\tilde{C}) \leqslant \lambda$ 的最大值 d 作为关联空间的秩. 例如假设 $k=2$, 不妨假设 $aveE(\tilde{C}) = 0.02$, 其中 $\frac{E(\tilde{C}_1, \varepsilon_1)}{E(\tilde{C}, \varepsilon_1)} = 0.02$, $\frac{E(\tilde{C}_2, \varepsilon_2)}{E(\tilde{C}, \varepsilon_2)} = 0.02$. 由定义 7 可知, 在 ε_1 下, $\tilde{C}_1$ 中的癌症样本到其质心的距离 $\tilde{C}$ 是中的癌症样本到其质心的距离的 1/50, 所以可以从 $\tilde{C}$ 中识别癌症 $\tilde{C}_1$. 同理, 在 ε_2 下可以从 $\tilde{C}$ 中识别癌症 $\tilde{C}_2$.

3.6 分类模型

假设已知癌症样本训练集(Training Set) 和测试集(Test Set), 训练集中样本的癌症类型都已知, 而测试集中样本的癌症类型未知. 如何根据关联空间来判断测试集中样本的癌症类型, 本小节提出一种基于权重的关联空间的分类模型(Weight based Classification with Related Space, WCRS). WCRS 的主要思想是将训练集中每个样本 $\bar{s}_l$ 与测试样本 $\bar{t}$ 在 $\bar{s}_l$ 所属癌症类型 $\tilde{C}_i$ 的最小扩展空间 ε_i 上的映射距离 $D(\bar{s}_l, \bar{t}, \varepsilon_i)$ 作为 $\bar{s}_l$ 在 ε_i 上对 $\bar{t}$ 的权重, 然后将训练集中癌症类型 $\tilde{C}_i (1 \leqslant i \leqslant k)$ 的所有样本权重之和作为 $\bar{t}$ 属于 $\tilde{C}_i$ 的权重, 即 $\sum_{\bar{s}_l \in \tilde{C}_i} D(\bar{s}_l, \bar{t}, \varepsilon_i)$, 最后取具有最小权重的癌症 $\tilde{C}_i'$ 为 $\bar{t}$ 的癌症类型.

WCRS 可以用如下数学模型来描述:

$$preV(\bar{t}) = \sum_{l, \bar{s}_l \in training set} weight(\bar{s}_l, \bar{t})\, vec(\bar{s}_l)$$

$$pre(\bar{t}) = classfmin(preV(\bar{t})) \quad (5)$$

其中 $\bar{s}_l \in \tilde{C}_i$, $weight(\bar{s}_l, \bar{t}) = D(\bar{s}_l, \bar{t}, \varepsilon_i)$, ε_i 是 $\tilde{C}_i$ 的最小扩展空间, $vec(\bar{s}_l)$ 是 $\bar{s}_l$ 的癌症类型矢量. 设存在三种类型癌症 $\tilde{C}_1$, $\tilde{C}_2$, 和 $\tilde{C}_3$, 其中 $\bar{s}_l \in \tilde{C}_2$, 则 $vec(\bar{s}_l) = (0, 1, 0)$. $classfmin(preV(\bar{t}))$. 表示 $preV(\bar{t})$ 中最小元素对应的癌症类型.

4 算法分析

由组能量的定义可知, 癌症 $\tilde{C}_i$ 在关联空间 ε_i 的组能量反映的是 $\tilde{C}_i$ 的癌症样本和 $\tilde{C}_i$ 的质心在 ε_i 上映射后的距离, 是 $\tilde{C}_i$ 中的癌症样本在 ε_i 上相似性的度量, 组能量越小, 组内样本的相似程度越大, 反之组内样本的相异程度越大. 我们利用组能量来评价癌症 $\tilde{C}_i$ 的关联空间 ε_i. 考虑本文提出的关联空间, 有如下定理:

定理 3　对于癌症样本集合 $\tilde{C}_i$ 和表达矩阵 $\boldsymbol{C}_i$, 假设 ε_i 是 $\tilde{C}_i$ 的关联空间, ε_i 的秩为 d, $\lambda_r (1 \leqslant r \leqslant d)$ 是 ε_i 的方向扩展系数, 则 $E(\tilde{C}_i, \varepsilon_i) = \sum_{i=1}^{d} \lambda_r$.

证明　对于癌症样本集合 $\tilde{C}_i$, 由于

$$E(\tilde{C}_i, \varepsilon_i) = \frac{1}{n_i} \sum_{l=1}^{n_i} \sum_{r=1}^{d} (\bar{s}_l \cdot \bar{p}_r - M(\tilde{C}_i) \cdot \bar{p}_r)^2 = \frac{1}{n_i} \sum_{r=1}^{d} \sum_{l=1}^{n_i} (\bar{s}_l \cdot \bar{p}_r - M(\tilde{C}_i) \cdot \bar{p}_r)^2,$$

又由定理 1 知, $\sum_{l=1}^{n_i} (\bar{s}_l \cdot \bar{p}_r - M(\tilde{C}_i) \cdot \bar{p}_r)^2 = n_i \lambda_r$, 因此 $E(\tilde{C}_i, \varepsilon_i) = \sum_{i=1}^{d} \lambda_r$. 证毕.

推论　设癌症样本集合 $\tilde{C}_i$ 和表达矩阵 $\boldsymbol{C}_i$, 如果 $\varepsilon_i = \{\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_m\}$, 则 $E(\tilde{C}_i, \varepsilon_i) = \mathrm{TR}(Cov(\boldsymbol{C}_i^{\mathrm{T}}))$.

证明　设 $\tilde{C}_i = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_{n_i}\}$, $\bar{s}_l = (\bar{s}_{l1}, \bar{s}_{l2}, \cdots, \bar{s}_{lm})^{\mathrm{T}}$, s_{lj}, 是基因 $\bar{g}_j$ 在 $\bar{s}_l$ 中的表达. 如果 $\varepsilon_i = \{\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_n\}$, 则

$$E(\tilde{C}_i, \varepsilon_i) = E(\tilde{C}_i, \{\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_m\}) = \frac{1}{n_i} \sum_{l=1}^{n_i} \| \bar{s}_l - M(\tilde{C}_i) \|.$$

又 $M(\tilde{C}_i) = \frac{1}{n_i} \sum_{l=1}^{n_i} \bar{s}_l$, 则

$$E(\tilde{C}_i, \varepsilon_i) = \frac{1}{n_i} \sum_{l=1}^{n_i} \sum_{j=1}^{m} \left(s_{lj} - \frac{1}{n_i} \sum_{l=1}^{n_i} s_{lj} \right)^2.$$

由定义(2) 可知, $\mathrm{TR}(Cov(\boldsymbol{C}_i^{\mathrm{T}})) = \sum_{j=1}^{m} Var(\bar{g}_j)$. 因为 $Var(\bar{g}_j) = \frac{1}{n_i} \sum_{k=1}^{n_i} \left(g_{jk} - \frac{1}{n_i} \sum_{j=1}^{n_i} g_{jk} \right)^2$, 且 $s_{lj} = g_{ji}$, 所以

$$\mathrm{TR}(Cov(C_i^{\mathrm{T}})) = \sum_{j=1}^{m} Var(\bar{g}_j) = \frac{1}{n_i} \sum_{l=1}^{n_i} \sum_{j=1}^{m} \left(s_{lj} - \frac{1}{n_i} \sum_{l=1}^{n_i} s_{lj} \right)^2.$$

因此, $E(\tilde{C}_i, \varepsilon_i) = \mathrm{TR}(Cov\boldsymbol{C}_i^{\mathrm{T}}))$. 证毕.

由推论可知, 癌症的关联空间的秩等于癌症样本的维数时, 即 $d = m$, 癌症 $\tilde{C}_i$ 具有最大的组能量, 并且等于没有进行关联空间映射时的组能量. 由定理 3 知, 癌症 $\tilde{C}_i$ 在最小扩展空间 ε_i 上的组能量小于在秩相同的关联空间 ε_i 上的组能量, 即在最小扩展空间 ε_i 上癌症 $\tilde{C}_i$ 中的样本比在秩相同的关联空间 ε_i 上具有更大的相似性.

5 实验结果分析

5.1 噪声基因的过滤

利用文献[12] 的"分类信息指数"(Information Index to Classification, IIC) 来度量基因包含样本分类信息量, 即 $d(\bar{g}) = \frac{1}{2} \frac{|\mu_{\bar{g}+} - \mu_{\bar{g}-}|}{\sigma_{\bar{g}+} - \sigma_{\bar{g}-}} + \frac{1}{2} \ln \left[\frac{\sigma_{\bar{g}+}^2 + \sigma_{\bar{g}-}^2}{2\sigma_{\bar{g}+}\sigma_{\bar{g}-}} \right]$. 其中, $\mu_{\bar{g}+}$, $\mu_{\bar{g}-}$ 分别为基因 $\bar{g}$ 在两个类别中表达水平的均值, $\sigma_{\bar{g}+}$, $\sigma_{\bar{g}-}$ 为对应的标准差.

5.2 癌症关联空间分析

分析对象为 Golub 等公布的急性白血病基因表达谱数据集(Leukemia)[3]. 该数据集共有 72 例急性白血病

样本, 每个样本均含 7129 个基因的表达数据. 其中 47 例样本被诊断为急性淋巴性白血病(Acute Lymphoblastic Leukemia, ALL), 25 例被诊断为急性骨髓性白血病(Acute Myeloid Leukemia, AML).

选取分类信息指数大于 0.8 的 196 个特征基因. 在急性白血病数据集 $\tilde{C}$ 中, 分别提取 47 例 ALL 的关联空间 ε_{ALL}和 25 例 AML 的关联空间 ε_{AML}, ε_{ALL}和 ε_{AML}的秩 d = 196. 图 2 给出了不同秩的最小扩展空间$(\hat{\varepsilon}_{ALL}, \hat{\varepsilon}_{AML})$下 $\tilde{C}$ 的组能量均值$averE(\tilde{C})$的情况.

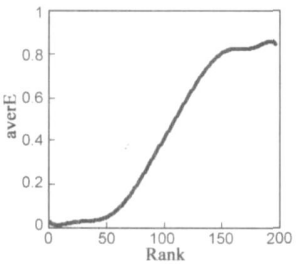

图 2 组能量均值随秩 (Rank) 的变化情况

当 λ 取 0.02 时, $\hat{\varepsilon}_{ALL}$和 $\hat{\varepsilon}_{AML}$的秩 $\hat{d}$ = 20. 图 3 和图 4 分别给出了 ALL 和 AML 在 $\hat{\varepsilon}_{ALL}$、$\hat{\varepsilon}_{AML}$下的分布情况. 我们分别将 $\hat{\varepsilon}_{ALL}$和 $\hat{\varepsilon}_{AML}$中的方向按其扩展系数进行递增排序, 称 $\hat{\varepsilon}_{ALL}$和 $\hat{\varepsilon}_{AML}$对应的方向为 $dim1$, $dim2$, …, $dim20$. 图 3 中(a)选取的方向是 ε_{ALL}的 $dim1$~ $dim3$, (b)选取的是 $dim5$~ $dim7$, (c)选取的是 $dim8$~ $dim10$, (d)选取的是 $dim11$~ $dim13$, (e)选取的是 $dim15$~ $dim17$, (f)选取的方向是 $dim18$~ $dim20$. 图 4 中选取 $\hat{\varepsilon}_{AML}$方向的序号与图 3 相同. 在 $\hat{\varepsilon}_{ALL}$下, 可以发现癌症模式 ALL, 图 3(a)~(f)中所有的 ALL 样本分布都非常集中, 虽然图(f)中有一个 AML 样本混淆在癌症 ALL 中, 图(d)中癌症 ALL 和癌症 AML 的界线比较模糊, 但是图(a)、(b)、(c)都可以有效的识别 ALL. 同样在 $\hat{\varepsilon}_{AML}$下也可以发现癌症模式 AML, 图 4(a)~(f)中所有的 AML 样本分布都非常集中, 虽然在图(a)、(b)、(d)中癌症 ALL 和癌症 AML 的界线比较模糊, 但在图(c)、(e)、(f)中可以有效地识别 AML.

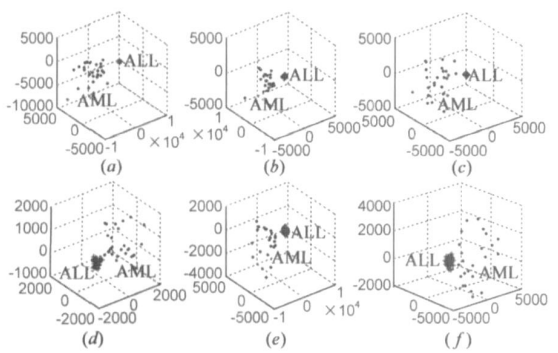

图 3 ALL 和 AML 在 $\hat{\varepsilon}_{ALL}$下的分布

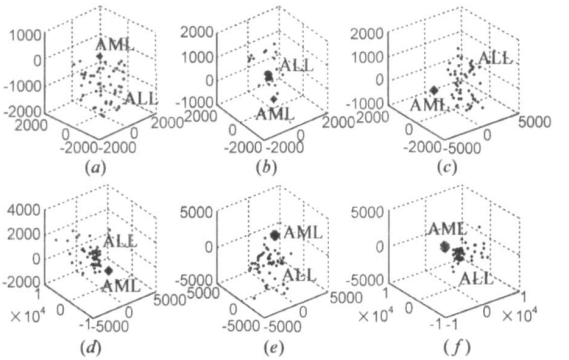

图 4 ALL 和 AML 在 $\hat{\varepsilon}_{AML}$下的分布

5.3 癌症分类结果分析

分别在急性白血病数据集和结肠癌数据集(Colon)上进行癌症识别. 结肠癌数据集共有 62 例样本, 其中包括 40 例结肠癌组织和 22 例正常结肠组织, 每例样本均含 2000 个基因的表达数据.

采用“留一法”(Leave One Out Cross Validation, LOOCV)[13]进行样本的癌症类型识别, 即在样本集上每次保留一个不同的样本作为测试样本, 其余样本作为训练数据集. 重复该过程, 直到每一个样本都有一次作为测试样本时为止. 统计所有被正确识别的样本数作为“留一法”的识别正确数. 表 1 和表 2 给出了最小扩展空间中的基因特征数目及其 WCRS 的分类性能, 并与 Golub 提出的加权投票法(Weighted voting)、Conde 提出的基于聚类(Clustering)的感知器模型、SVM 和 KNN 的分类结果进行了比较. 在 WCRS 中选取分类信息指数阈值取 0.8, 在提取最小扩展空间时, 参数取 0.02. 在 Leukemia 和 Colon 中特征基因数目分别为 20 和 22. 根据经验值选取其余方法中分类较优的特征基因数目, 且与 WCRS 中的基因数目接近, 在 Leukemia 中选取 50 个特征基因, 在 Colon 中选择 55 个特征基因. SVM 采用径向基函数(RBF)作为核函数, KNN 相似性度量函数采用 Pearson 相关系数. 从表 1 和表 2 可以看出, 本文方法提取的最小扩展空间方法更加有效地压缩了基因表达数据维数, WCRS 提高了样本的识别正确率.

表 1 在白血病数据集上的实验结果比较

分类方法	特征基因数目	识别正确数
WCRS	20	71
Weighted voting	50	65
SVM	50	69
KNN	50	62
Clustering	50	63

表 2 在结肠癌数据集上的实验结果比较

分类方法	基因特征数目	识别正确数
WCRS	22	60
Weighted voting	55	57
SVM	55	58
KNN	55	51
Clustering	35	53

6 结论

针对基因表达谱数据中致癌因子存在局部相关性的特点,提出一种基于权重的关联空间的癌症分类算法.首先构造癌症组的关联空间,并提出基于关联空间的基因表达模式,然后提取使得癌症组具有最小组能量的最小扩展空间,最后在最小扩展空间上构建一种基于权重的癌症分类模型.该算法不仅有效地压缩了基因表达数据维数,并比传统分类算法具有更好的分类精度.如何更好地获取关联空间的维数,是我们下阶段研究的重点.

参考文献:

[1] M Kuramochi, G Karypis. Gene classification using expression profiles: a feasibility study[J]. International Journal on Artificial Intelligence Tools, 2005, 14(4): 641- 660.

[2] 李霞,张田文,郭政. 一种基于递归分类树的集成特征基因选择方法[J]. 计算机学报, 2004, 27(5): 675- 682.
Li Xia, Zhang Tianr wen, Guo Zheng. An novel ensemble method of feature gene selection based on recursive partitionr tree[J]. Chinese Journal of Computers, 2004, 27(5): 675- 682. (in Chinese)

[3] M A Shipp, K N Ross, P Tamayo, et al. Diffuse large B-cell lymphoma outcome prediction by gene- expression profiling and supervised machine learning[J]. Nature Medicine, 2002, 8(1): 68- 74.

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

[5] L J V T Veer, H Dai, M J V D Vijver, Y D He, et al. Gene expression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871): 530- 536.

[6] S B Cho, H H Won. Machine learning in DNA Microarray analysis for cancer classification[A]. First Asiar Pacific Bioinformatics Conference(APBC 2003)[C]. Adelaide, Australia: Australian Computer Society, 2003. 189- 198.

[7] X G Lu, Y P Lin, X L Yang, et al. Using most similarity tree based clustering to select the top most discriminating genes for cancer detection[A]. 8th International Conference of Artificial Intelligence and Soft Computing(ICAISC 2006)[C]. Berlin, German: Springer-Verlag, 2006. 931- 940.

[8] X Hu, I Yoo. Cluster ensemble and its application in gene expression analysis[A]. 2nd Asiar Pacific Bioinformatics Conference(APBC2004)[C]. Adelaide, Australia: Australian Computer Society, 2004. 297- 302.

[9] L Conde, A Mateos, J Herrero, et al. Unsupervised reduction of the dimensionality followed by supervised learning with a perceptron improves the classification of conditions in DNA Microarray gene expression data[A]. Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing[C]. 2004, 77- 86.

[10] J Khan, J S Wei, M Ringne, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artifr cial neural networks[J]. Nature Medicine, 2001, 7(6): 673- 679.

[11] L Parsons, E Haque, H Liu. Subspace clustering for high dimensional data: a review [J]. SIGKDD Explorations, 2004, 6(1): 90- 105.

[12] 李颖新,阮晓钢. 基于基因表达谱的肿瘤亚型识别与分类特征基因选取研究[J]. 电子学报, 2005, 33(4): 651- 655.
Li Ying-xin, Ruan Xiaor gang. Cancer subtype recognition and feature selection with gene expression profiles. Acta Electronica Sinica[J]. 2005, 33(4): 651- 655. (in Chinese)

[13] J Y Li, L Wong. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns [J]. Bioinformatics, 2002, 18(5): 725- 734.

作者简介:

卢新国　男, 1979年生于湖南汨罗, 2002年获湖南大学数学与计量经济学院工学学士学位,湖南大学计算机与通信学院博士,主要研究方向为模式识别、数据挖掘与机器学习、生物信息学. E-mail: hnluxinguo@hotmail.com

林亚平　男, 1955年生于湖南邵阳, 1982年获湖南大学工学学士学位, 1985年获国防科技大学工学硕士学位, 2000年获湖南大学工学博士学位.现任湖南大学件学院和计算机与通信学院教授,博士生导师.主要研究领域为计算机网络、数据挖掘与机器学习、生物信息学.
E-mail: yplin@hnu.cn