基于类别保留投影的基因表达数据特征提取新方法

王文俊
(西安电子科技大学计算机学院,陕西西安 710071)

摘 要: 从两两样本的类别关系出发,提出一种新的线性鉴别特征提取方法,叫做类别保留投影.相比经典的fisher线性鉴别分析方法,类别保留投影具有最优子空间维数不受样本类别数限制、计算复杂度低的优点.通过对真实基因表达数据进行样本分类识别,证实了本文方法的有效性.并将类别保留投影方法推广到非线性空间,提出核类别保留投影,用于解决非线性特征提取问题,对基因表达数据的实验验证了方法的可行性.

关键词: 特征提取; fisher线性鉴别分析; 小样本; 基因表达数据
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2012) 02-0358-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.02.024

New Method of Feature Extraction for Gene Expression Data Based on Class Preserving Projection

WANG Wen-jun
(*School of Computer Science and Engineering, Xidian University. Xi'an, Shaanxi* 710071, *China*)

Abstract: A new method of discriminant feature extraction, called Class Preserving Projection (CPP), is proposed from the point view of class relation of pairwise samples. Compared to LDA, CPP has the following two advantages. One is that the optimal subspace dimension is not restricted to the number of categories of samples, and the other is that computational complexity is lower. Experiments are performed on gene expression data for sample classification, and the results confirm the effectiveness of the method. Kernel CPP (KCPP) is presented by generalizing CPP to nonlinear space to solve the problem of nonlinear feature extraction, and the experiments on gene expression data verify the feasibility of the method.

Key words: feature extraction; Fisher's linear discriminant analysis(LDA); small sample size; gene expression data

1 引言

DNA微阵列技术[1,2]能同时检测成千上万个基因的表达水平,为人类在分子水平进行疾病的诊断和治疗提供了全新的手段.如何利用基因表达数据[3]进行疾病的诊断和治疗,就成为生物信息领域研究的热点问题.模式分类[4]是信息挖掘的关键技术之一,常用于组织样本的分类识别,如癌症分型[5].特征提取是实现样本分类前数据降维的有效方法之一,经典且常用的特征提取方法之一就是fisher线性鉴别分析(LDA)[6~9].LDA利用样本的已知类别信息指导特征的学习,通过最大化类间散度而最小化类内散度来寻找最优鉴别特征.然而,利用LDA进行基因表达数据特征提取时,存在如下两个问题:一是LDA的最优子空间维数受样本类别数 c 限制,最优子空间的维数不超过 $c-1$;二是基因表达数据的协方差矩阵的计算复杂度高,随着基因数的增加呈指数增长.

为了有效提取基因表达数据的鉴别信息,同时克服LDA的以上两个缺点,本文从两两样本的类别关系出发,提出一种新的鉴别特征提取方法——类别保留投影(Class Preserving Projection, CPP).通过最小化类内散度而最大化类间散度来构造目标函数,把两两样本的类别关系作为权重系数,使得同类样本尽可能地聚集,而异类样本尽可能地发散,来寻找线性最优鉴别特征.与LDA相比,CPP能获得更高维的最优子空间,不需要计算协方差矩阵,能有效降低特征提取的计算复杂度.对于非线性问题,本文通过"核技巧"将CPP推广到非线性空间,提出核类别保留投影(Kernel Class Preserving Projection, KCPP).通过对真实基因表达数据的实验,验证了CPP和KCPP用于基因表达数据特征提取的有效性和可行性.

首先,我们将线性投影问题描述如下:

设 $n \times m$ 维的基因表达数据矩阵 $\boldsymbol{X}$,行代表基因,列代表组织样本(简称"样本"),其元素 $\boldsymbol{x}_{ij}$ 是基因 i 在样

收稿日期:2011-02-16;修回日期:2011-09-16
基金项目:中央高校基本科研业务费专项资金(No.K5051203013)

本 j 上的表达水平.每个样本对应一个 n 维的表达向量,即 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\in R^n$.找一个变换矩阵 $\boldsymbol{A}$,使这 m 个样本映射到 l 维空间中的 m 个点:$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_m\in\boldsymbol{R}^l$,使得 $\boldsymbol{y}_i$ 代表 $\boldsymbol{x}_i$,这里 $\boldsymbol{y}_i=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i$.

2 LDA

设 m 个样本属于 c 类,第 i 类的样本数是 m_i,且有 $\sum_{i=1}^{c}m_i=m$. $\boldsymbol{x}_i^j,i=1,\cdots,m_j,j=1,\cdots,c$ 表示第 j 类的第 i 个样本.设 $\boldsymbol{a}$ 是一个最优变换向量,LDA 的目标函数是:

$$\max\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{a}} \tag{1}$$

其中,矩阵 $\boldsymbol{S}_b$ 和 $\boldsymbol{S}_w$ 分别是类间散布矩阵和类内散布矩阵,定义如下:

$$\boldsymbol{S}_b=\frac{1}{m}\sum_{j=1}^{c}m_j(\boldsymbol{\mu}_j-\boldsymbol{\mu})(\boldsymbol{\mu}_j-\boldsymbol{\mu})^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{S}_w=\frac{1}{m}\sum_{j=1}^{c}\sum_{i=1}^{m_j}(\boldsymbol{x}_i^j-\boldsymbol{\mu}_j)(\boldsymbol{x}_i^j-\boldsymbol{\mu}_j)^{\mathrm{T}} \tag{3}$$

其中,$\boldsymbol{\mu}$ 是所有样本的均值,$\boldsymbol{\mu}_j$ 是第 j 类样本的均值.

最优变换向量 $\boldsymbol{a}$ 可通过解下列广义特征方程获得:

$$\boldsymbol{S}_b\boldsymbol{a}=\lambda\boldsymbol{S}_w\boldsymbol{a} \tag{4}$$

式(4)的最大特征值对应的特征向量就是 LDA 的最优变换向量 $\boldsymbol{a}$.

矩阵 $\boldsymbol{S}_w$ 和 $\boldsymbol{S}_b$ 的维数都是 $n\times n$,故直接利用式(4)求解,其计算复杂度与基因数 n 有关,是基于基因空间的求解.基因数 n 越大,则 LDA 求解过程中特征方程求解的时间就越长.同时对于基因表达数据来说,由于 $n\gg m$,所以 $\boldsymbol{S}_w$ 是严重奇异的.为了降低矩阵 $\boldsymbol{S}_w$ 的奇异程度、同时降低广义特征方程的计算复杂度,我们通过简单的代数变换,把 LDA 的求解过程从基因空间转换到样本空间,具体转换方法如下:

LDA 是利用 m 个训练样本寻找一个最优线性变换,任一线性变换向量必位于所有训练样本在原属性空间的张集,可找到最优变换向量 $\boldsymbol{a}$ 的一种展开式:

$$\boldsymbol{a}=\sum_{i=1}^{m}\beta_i\boldsymbol{x}_i=\boldsymbol{X}\boldsymbol{\beta} \tag{5}$$

其中,$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m]$,$\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots,\beta_m]^{\mathrm{T}}$ 是变换向量 $\boldsymbol{a}$ 在各个样本上的权重系数.将式(5)代入式(1),LDA 的目标函数就变为

$$\max\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{a}}\Leftrightarrow\max\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{X}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}\boldsymbol{\beta}} \tag{6}$$

使式(6)的目标函数最大的系数向量 $\boldsymbol{\beta}$ 可以通过求解以下的广义特征方程来获得:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{X}\boldsymbol{\beta}=\lambda'\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}\boldsymbol{\beta} \tag{7}$$

广义特征方程(7)的最大特征值对应的特征向量就是最优变换向量 $\boldsymbol{a}$ 的系数向量 $\boldsymbol{\beta}$,由 $\boldsymbol{\beta}$ 通过式(5)可求得最优变换向量为 $\boldsymbol{a}$.

矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{X}$ 的维数都是 $m\times m$ 的,m 是样本个数,广义特征方程(7)的求解空间转换到了样本空间,所以称以上 LDA 的求解方法为基于样本空间的 LDA.在样本空间,数据的类内散布矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}$ 的维数是 $m\times m$,而由 m 个样本计算的矩阵的秩不超过 $m-1$,所以矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}$ 依然是奇异的,不过相比基因空间 $n\times n$ 维的类内散布矩阵 $\boldsymbol{S}_w$,其奇异程度已大大降低.可采用现有的解决 LDA 矩阵奇异问题的方法解决矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}$ 的奇异问题.本文主要采用三种方法:一是用矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}$ 的伪逆矩阵代替其逆矩阵的计算("伪逆代替逆")[10];二是采用 PCA 的方法对矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{X}$ 进行K-L展开,丢掉其零空间信息,再在其非零空间进行 LDA 特征提取[11];三是采用 PCA 的方法对矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{X}$ 进行K-L展开,丢掉其零空间信息,再在其非零空间进行 LDA 特征提取[12].

LDA 是由 c 个类均值计算的协方差矩阵作为类间散布矩阵 $\boldsymbol{S}_b$,而由 c 个样本所撑子空间的维数不超过 $c-1$,所以 LDA 的最优子空间的维数最大不超 $c-1$.对于基因表达数据的样本特征提取来说,样本的类别数 c 都很少(如 2 类到 5 类),用 LDA 进行特征提取的最优子空间维数就很少.同时,LDA 通过计算类内协方差矩阵和类间协方差矩阵分别获得类内散布矩阵和类间散布矩阵,对于基因表达数据来说,协方差矩阵的维数与基因数 n 有关,是 $n\times n$ 维的,所以协方差矩阵的计算复杂度将会随着基因数的增加呈指数增长.

3 CPP

CPP 从两两样本的类别关系出发,样本的类别关系作为权重系数,构造目标函数,使同类的任意两样本的距离尽可能地小,而异类的任意两样本之间的距离尽可能地大.

设 $\boldsymbol{a}$ 是一个变换向量,样本 $\boldsymbol{x}_i$ 在 $\boldsymbol{a}$ 上的投影记为 y_i,即 $y_i=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i,i=1,\cdots,m$,样本 x_i 的类别记为 c_i.则可构造 CPP 的目标函数为

$$\min\frac{\sum_{ij}(y_i-y_j)^2\boldsymbol{W}_{ij}^1}{\sum_{ij}(y_i-y_j)^2\boldsymbol{W}_{ij}^2} \tag{8}$$

其中,

$$\boldsymbol{W}_{ij}^1=\begin{cases}1, & \text{if}\quad c_i=c_j;\\ 0, & \text{else.}\end{cases} \tag{9}$$

$$\boldsymbol{W}_{ij}^2=\begin{cases}1, & \text{if}\quad c_i\neq c_j;\\ 0, & \text{else.}\end{cases} \tag{10}$$

权重系数矩阵 $\boldsymbol{W}^1$ 使得目标函数的分子只计算同类两两样本的距离,而权重系数矩阵 $\boldsymbol{W}^2$ 使得目标函数的分母只计算不同类两两样本的距离.通过最小化目

标函数,就使得同类任意两样本之间的平均距离达到最小,而不同类任意两样本之间的平均距离达到最大.

把 $y_i = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i, i = 1,\cdots,m$ 代入式(8),通过简单的代数变换,目标函数可化简为

$$\frac{\frac{1}{2}\sum_{ij}(y_i - y_j)^2 \boldsymbol{W}_{ij}^1}{\frac{1}{2}\sum_{ij}(y_i - y_j)^2 \boldsymbol{W}_{ij}^2} = \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}} \tag{11}$$

其中,$\boldsymbol{X}$ 的第 i 列是 $\boldsymbol{x}_i$,$\boldsymbol{D}^k, k = 1,2$ 是对角矩阵,其元素值是矩阵 $\boldsymbol{W}^k$ 的列(或行)元素之和,即

$$\boldsymbol{D}_{ii}^k = \sum_j \boldsymbol{W}_{ij}^k \tag{12}$$

且

$$\boldsymbol{L}^k = \boldsymbol{D}^k - \boldsymbol{W}^k, k = 1,2 \tag{13}$$

使目标函数(11)最小的变换向量 $\boldsymbol{a}$ 可通过求解以下的广义特征方程来获得:

$$\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} = \lambda\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} \tag{14}$$

可见,矩阵 $\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}$和 $\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}$是对称且半正定的.广义特征方程(14)的最小特征值对应的特征向量就是最优变换向量 $\boldsymbol{a}$.广义特征方程的前 l 个最小特征值对应的特征向量 $\boldsymbol{a}_i(i = 1,2,\cdots,l)$就构成了 CPP 的最优变换矩阵 $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_l)$.

从 CPP 的目标函数来看,矩阵 $\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}$是样本的类内散布矩阵,矩阵 $\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}$是样本的类间散布矩阵.最小化 CPP 的目标函数,就使得样本的类内散度最小而类间散度最大,从这点来看,CPP 与 LDA 是相同的.从类内散度和类间散度的计算方法来看:CPP 的类内散度是以类内两两样本的平均距离度量的,LDA 的类内散度以类条件方差来度量,二者也是一致的;但是 CPP 的类间散度是以类间两两样本的平均距离度量的,而 LDA 的类间散度是以类均值之间的平均距离度量的,二者是不同的.CPP 的类间散布矩阵是由 m 个训练样本计算的,所以其秩不超过 $m - 1$;而 LDA 的类间散布矩阵是由 c 个类均值计算的,所以其秩不超过 $c - 1$.正是由于类间散度的计算方法不同于 LDA,才使得 CPP 的最优子空间的维数不受样本类别数的限制.所以 CPP 能找到比 LDA 更高维的最优子空间.同时,CPP 的类内散布矩阵和类间散布矩阵是通过矩阵相乘获得,而不需要像 LDA 那样计算类内协方差矩阵和类间协方差矩阵,这将大大降低散布矩阵的计算复杂度,数据的维数 n 越大,CPP 的计算优势就越明显.

CPP 的类内散布矩阵 $\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}$和类间散布矩阵 $\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}$都是 $n \times n$ 维的,n 是数据的维数(基因数).所以广义特征方程(14)的计算是基于基因空间的,当基因数 n 很大时,广义特征方程(14)的计算复杂度将会很高.同时,对于基因表达数据而言,通常 $n >> m$,所以由 m 个样本计算的类间散布矩阵 $\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}$一定是严重奇异的.这与 LDA 特征提取方法中存在的问题是一样的.为解决这个问题,类似于 LDA,我们采用基于样本空间的 CPP.

CPP 是利用 m 个训练样本寻找一个线性变换,任一线性变换向量 $\boldsymbol{a}$ 必位于所有训练样本在原属性空间的张集,因此可找到最优变换向量 $\boldsymbol{a}$ 的一种展开式:

$$\boldsymbol{a} = \sum_{i=1}^{m} \beta_i \boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{\beta} \tag{15}$$

其中,$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m]$,$\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_m]^{\mathrm{T}}$是变换向量 $\boldsymbol{a}$ 在各个样本上的权重系数.

将式(15)代入式(11),CPP 的目标函数就变为

$$\min \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}} \Leftrightarrow \max \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}} \tag{16}$$

使目标函数式(16)最小的系数向量 $\boldsymbol{\beta}$ 可以通过求解以下的广义特征方程来获得:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \lambda'\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} \tag{17}$$

特征方程式(17)的最小特征值对应的特征向量就是最优变换向量 $\boldsymbol{a}$ 的系数向量$\boldsymbol{\beta}$,则最优变换向量 $\boldsymbol{a} = \boldsymbol{X}\boldsymbol{\beta}$.

由以上分析可知,矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的维数都是 $m \times m$ 的,m 是样本个数,广义特征方程(17)的求解空间转换到了样本空间,所以称以上 CPP 的求解方法为基于样本空间的 CPP.在样本空间,数据的类间散布矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的维数是 $m \times m$,而由 m 个样本计算的类间散布矩阵的秩不超过 $m - 1$,所以矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 依然是奇异的,不过相比基因空间 $n \times n$ 维的类间散布矩阵 $\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}$,其奇异程度已大大降低.本文主要采用以下两种方法解决矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的奇异问题:一是用矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的伪逆矩阵代替其逆矩阵的计算("伪逆代替逆");二是采用 PCA 的方法对矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 进行 K-L 展开,丢掉其零空间信息,再在其非零空间进行 CPP 特征提取.

4 KCPP

由于 CPP 是线性投影方法,对非线性可分的数据而言,采用 CPP 进行特征提取和降维的效果并不理想.为了解决非线性的问题,我们利用"核技巧"首先把数据非线性地映射到某个特征空间,然后在这个特征空间中进行 CPP 投影,这样就隐含地实现了原输入空间的非线性判别.

设 $\boldsymbol{\Phi}$ 是输入空间 $\boldsymbol{R}^n$ 到某个特征空间的非线性映射,即:$\boldsymbol{X} \rightarrow \boldsymbol{\Phi}(\boldsymbol{X})$,其中,$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m]$,$\boldsymbol{\Phi}(\boldsymbol{X}) = [\boldsymbol{\Phi}(\boldsymbol{x}_1), \boldsymbol{\Phi}(\boldsymbol{x}_2), \cdots, \boldsymbol{\Phi}(\boldsymbol{x}_m)]$.在核空间 CPP 的目标函数为:

$$\min \frac{\sum_{ij}(y_i^{\Phi} - y_j^{\Phi})^2 \boldsymbol{W}_{ij}^{1\Phi}}{\sum_{ij}(y_i^{\Phi} - y_j^{\Phi})^2 \boldsymbol{W}_{ij}^{2\Phi}} \tag{18}$$

其中,y_i^{φ} 是 $\boldsymbol{\Phi}(\boldsymbol{x}_i)$在投影向量 $\boldsymbol{a}^{\Phi}$ 上的一维表示,即

$$y_i^{\Phi} = (\boldsymbol{a}^{\Phi})^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_i) \tag{19}$$

显然,如果核空间的维数很高甚至是无穷维的,直接求解是不可能的.为此,我们使用核技巧,无需对数据进行明确的映射,而是寻找算法的一种表达式,其中只使用了样本在核空间的点积运算,只要能够有效地计算这些点积运算,就能够解决原始的问题.可通过使用 Mercer 核来实现:这些核函数 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$计算了某个特征空间的点积运算,即 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{\Phi}(\boldsymbol{x}_i),\boldsymbol{\Phi}(\boldsymbol{x}_j))$.常用的核函数有多项式核函数 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\boldsymbol{x}_j+1)^d$,RBF 核函数 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|2/(2\sigma^2))$等.

最优的变换向量 $\boldsymbol{a}^{\Phi}$ 通过求解一个特征值问题得到.根据再生核理论,任何 $\boldsymbol{a}^{\Phi}$ 必位于所有训练样本在核空间的张集,因此可找到下列形式的 $\boldsymbol{a}^{\Phi}$ 的一个展开式

$$\boldsymbol{a}^{\Phi} = \sum_{p=1}^{m}\beta_p\boldsymbol{\Phi}(\boldsymbol{x}_p) = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{\beta} \tag{20}$$

其中 $\boldsymbol{\Phi}(\boldsymbol{X})=[\boldsymbol{\Phi}(\boldsymbol{x}_1),\boldsymbol{\Phi}(\boldsymbol{x}_2),\cdots,\boldsymbol{\Phi}(\boldsymbol{x}_m)]$,$\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots,\beta_m]^{\mathrm{T}}$.

将式(19)和式(20)代入式(18),得

$$\frac{\frac{1}{2}\sum_{ij}(y_i^{\Phi}-y_j^{\Phi})^2\boldsymbol{W}_{ij}^{1\Phi}}{\frac{1}{2}\sum_{ij}(y_i^{\Phi}-y_j^{\Phi})^2\boldsymbol{W}_{ij}^{2\Phi}} = \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{1\Phi}\boldsymbol{K}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{2\Phi}\boldsymbol{K}\boldsymbol{\beta}} \tag{21}$$

其中,$\boldsymbol{K}=(\boldsymbol{\Phi}(\boldsymbol{X}))^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{X})$,$\boldsymbol{L}^{k\Phi}=\boldsymbol{D}^{k\Phi}-\boldsymbol{W}^{k\Phi}$,$\boldsymbol{D}^{k\Phi}=\mathrm{diag}\left[\sum_j\boldsymbol{W}_{1j}^{k\Phi},\sum_j\boldsymbol{W}_{2j}^{k\Phi},\cdots,\sum_j\boldsymbol{W}_{mj}^{k\Phi}\right]$,$k=1,2$. $\boldsymbol{W}^{k\Phi}$是类别关系矩阵,其元素值是刻画核空间中两点的类别关系,它与原空间的类别关系一样,即

$$\boldsymbol{W}_{ij}^{1\Phi}=\boldsymbol{W}_{ij}^{1}=\begin{cases}1, & \text{if}\quad c_i=c_j;\\ 0, & \text{else}.\end{cases} \tag{22}$$

$$\boldsymbol{W}_{ij}^{2\Phi}=\boldsymbol{W}_{ij}^{2}=\begin{cases}1, & \text{if}\quad c_i\neq c_j;\\ 0, & \text{else}.\end{cases} \tag{23}$$

根据式(21),式(18)的最小化问题就变为了

$$\mathrm{argmin}_{\beta}\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{1\Phi}\boldsymbol{K}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{2\Phi}\boldsymbol{K}\boldsymbol{\beta}} \tag{24}$$

从而系数向量 $\boldsymbol{\beta}$ 可通过求解下列特征值问题得到:

$$\boldsymbol{K}\boldsymbol{L}^{1\Phi}\boldsymbol{K}\boldsymbol{\beta}=\lambda\boldsymbol{K}\boldsymbol{L}^{2\Phi}\boldsymbol{K}\boldsymbol{\beta} \tag{25}$$

其中 $\boldsymbol{K}=(\boldsymbol{\Phi}(\boldsymbol{X}))^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{X})$,$\boldsymbol{L}^{k\Phi}=\boldsymbol{D}^{k\Phi}-\boldsymbol{W}^{k\Phi}$,$\boldsymbol{D}^{k\Phi}=\mathrm{diag}\left[\sum_j\boldsymbol{W}_{1j}^{k\Phi},\sum_j\boldsymbol{W}_{2j}^{k\Phi},\cdots,\sum_j\boldsymbol{W}_{mj}^{k\Phi}\right]$,$k=1,2$. $\boldsymbol{W}_{ij}^{k\Phi}$由式(22)和式(23)给出.

式(25)对应的特征方程的最小特征值对应的特征向量 $\boldsymbol{\beta}^j(j=1,2,\cdots,l)$就是使目标函数最大的核空间的投影向量的权系数,$\boldsymbol{\beta}^j$ 是 m 维的向量.此时投影向量 $\boldsymbol{a}^{\Phi}=\sum_{p=1}^{m}\beta_p^j\boldsymbol{\Phi}(\boldsymbol{x}_p)=\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{\beta}^j$,给定一个新的样本 $\boldsymbol{x}_i$,其在 $\boldsymbol{a}^{\Phi}$ 上的一维投影是

$$y_{ji}^{\Phi}=(\boldsymbol{a}^{\Phi})^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_i)=(\boldsymbol{\beta}^j)^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x}_i) \tag{26}$$

给定一个新的样本 $\boldsymbol{x}_i$,采用 KCPP 方法投影后的 l 维投影向量 $\boldsymbol{Y}_i^{\Phi}=[y_{1i}^{\Phi},y_{2i}^{\Phi},\cdots,y_{li}^{\Phi}]^{\mathrm{T}}$,即

$$\boldsymbol{Y}_i^{\Phi}=[y_{1i}^{\Phi},y_{2i}^{\Phi},\cdots,y_{li}^{\Phi}]^{\mathrm{T}}=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x}_i) \tag{27}$$

其中,$\boldsymbol{\beta}$ 是由广义特征方程式(25)的前 l 个最小的特征值对应的特征向量组成的 $m\times l$ 的矩阵,即 $\boldsymbol{\beta}=[\boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\cdots,\boldsymbol{\beta}^l]$,且 $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x}_i)=(\boldsymbol{\Phi}(\boldsymbol{X}))^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_i)$.

5　实验结果

本节对本文提出的方法进行验证,并与 LDA 进行比较.给出 8 个真实基因表达数据的实验结果,分别是 SRBCT[13]、Brain[14]、Colon[15]、Lymphoma[16]、ALLAML[17]、Lung Cancer[18]、Prostate[19]和 GCM[20],数据具体描述见表 1.以下实验均是在 Intel 的 2GHz、4 核的处理器上,内存为 2GHz、6GB 的环境下运行的.由于基因表达数据的基因数远大于样本数,所以本节我们对 CPP 和 LDA 均采用基于样本空间的特征提取方法进行计算.对于基因表达数据来说,基于样本空间的 LDA 和 CPP 依然存在矩阵奇异问题.对 LDA,可采用第 2 节提到的三种不同的解决奇异的方法来进行 LDA 特征提取;对 CPP,可采用第 3 节提到的两种不同的解决奇异的方法进行 CPP 特征提取.

5.1　CPP 的实验结果

本节主要对 CPP 的计算复杂度、提取更高维最优子空间的有效性和特征的鉴别性能进行验证,并与 LDA 进行比较.

5.1.1　CPP 的计算复杂度验证

为比较 CPP 和 LDA 的计算复杂度,本节均采用"伪逆代替逆"的方法进行 CPP 和 LDA 的特征提取,其计算时间见表 1.

表 1　数据和基于样本空间的特征提取时间

数据	基因数(个)	样本数(个)	类数(个)	LDA 特征提取时间(秒)	CPP 特征提取时间(秒)
SRBCT	2308	64	4	1.4	0.62
Brain	5597	42	5	9.53	1.47
Colon	2000	62	2	2.73	10.2
Lymphoma	4026	62	3	3.83	2.54
ALLAML	7129	72	2	6.59	3.77
Lung Cancer	12533	181	2	642.4	16.8
Prostate	12600	102	2	1136	9.67
GCM	16063	190	14	102938	25.8

从表 1 中可看出,CPP 在所有数据上(除数据 Colon 外)的特征提取时间要明显少于样本空间 LDA 的特征提取时间.尤其当基因个数超过 1 万时,CPP 的计算优势就愈发明显.如基因数为 12533 的 Lung Cancer 数据,

其基于样本空间的 LDA 特征提取时间是 642.4 秒,而基于样本空间的 CPP 特征提取只要 16.8 秒.这是因 CPP 不需要计算协方差矩阵,所以相比 LDA,其计算时间要少.尤其随着基因数的增加,CPP 的计算优势就越明显,如对基因数为 16063 个的 GCM 数据进行基于样本空间的 CPP 特征提取,在当前运行环境下(内存为 6GB、虚拟内存为 4089MB)的计算空间就够了,而且所花的时间仅为 25.8 秒钟,相比样本空间 LDA 所花的 102938 秒(28.59 小时),其特征提取速度提高了近 4000 倍.

综上可知,本文提出的鉴别特征提取方法 CPP 相比 LDA 来说,有效地减低了特征提取的计算复杂度.那么,同样作为鉴别特征提取方法,CPP 的鉴别性能如何呢?5.1.2 节就主要分析和比较 CPP 与 LDA 的鉴别能力.

5.1.2 CPP 的鉴别性能验证

我们从学习能力和推广能力两方面进行验证.

(1)学习能力 首先我们对样本进行二维可视化,数据分别在 LDA 和 CPP 的前两个主分量上的散布见图 1 和图 2.从图 1 和图 2 中可看出,CPP 的二维可视化中,除两类数据外,对其他数据来说,同类样本几乎聚为了一点,而且不同类的样本分的比较开;而两类样本的数据的 CPP 二维可视化中,有一类样本几乎聚为了一点,而另一类数据在第一个主分量上也聚为了一点.这说明:相比 LDA,CPP 能保留更多的类别信息,其学习能力更强.

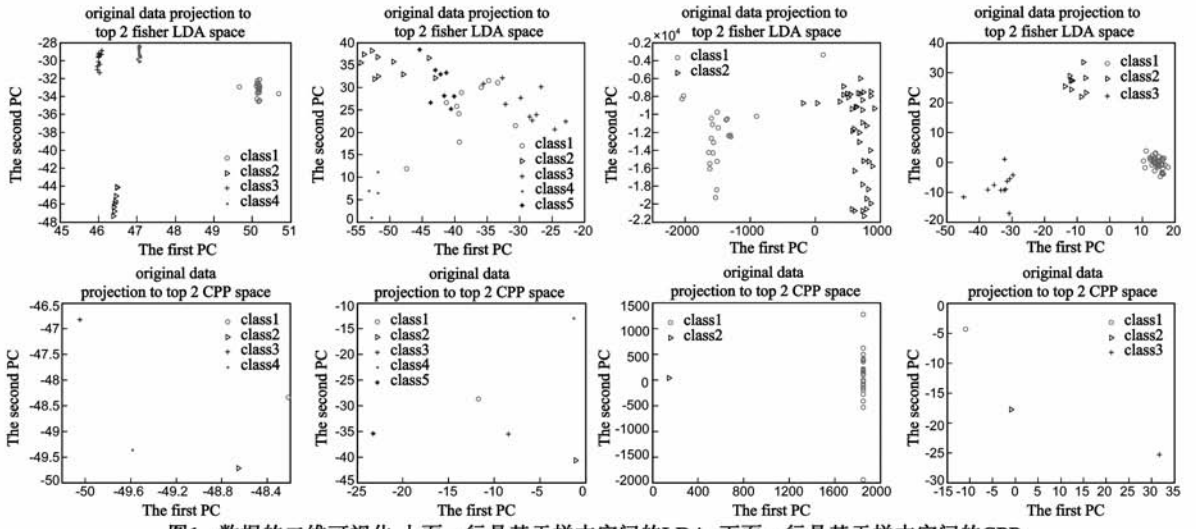


图1 数据的二维可视化.上面一行是基于样本空间的LDA;下面一行是基于样本空间的CPP.
四列从左到右依次对应SRBCT、Brain、Colon、Lymphoma数据.

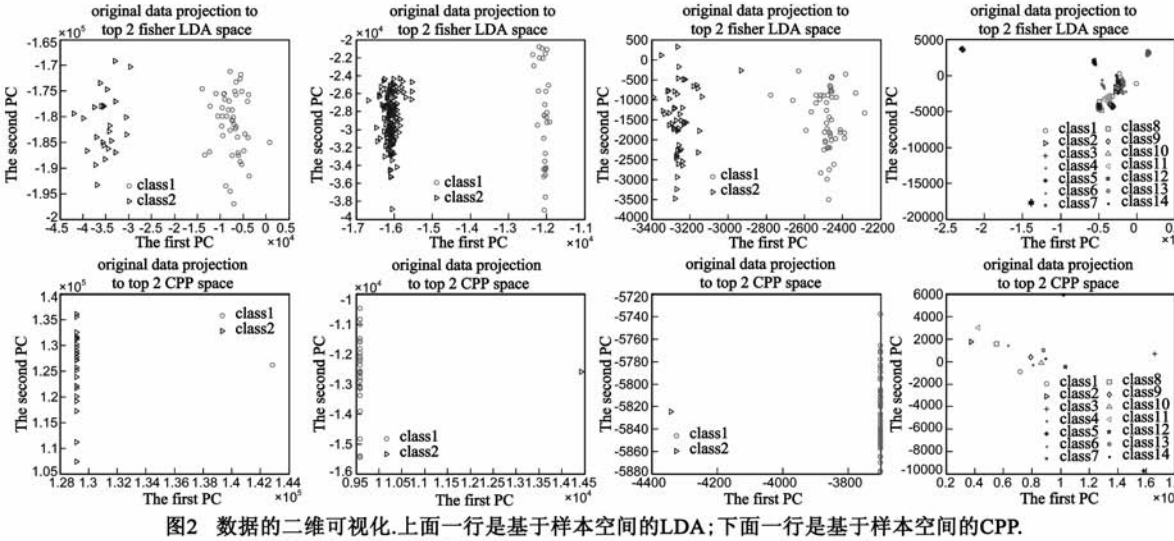
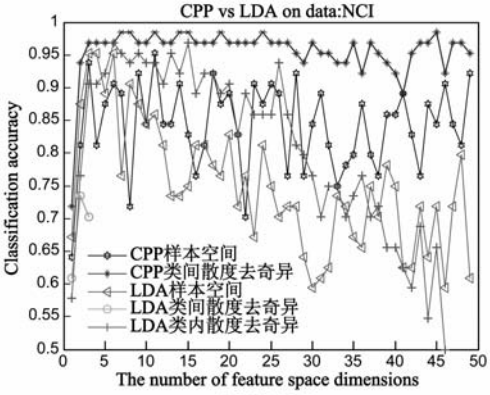


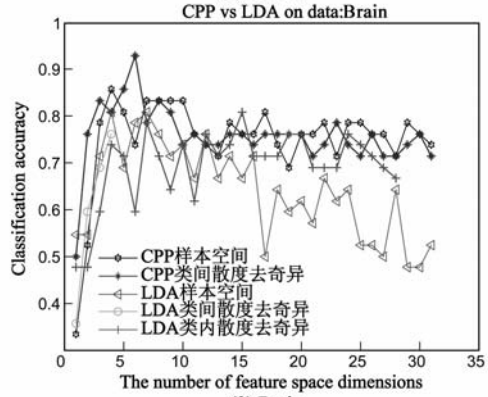
图2 数据的二维可视化.上面一行是基于样本空间的LDA;下面一行是基于样本空间的CPP.
四列从左到右依次对应ALLAML、Lung Cancer、Prostate、GCM数据.

(2)推广能力 下面我们验证 CPP 的推广能力.LDA 分别采用第 2 节提到的三种解决矩阵奇异的方法,CPP 分别采用第 3 节提到的两种解决奇异的方法.采用 5 重交叉验证来分析特征的推广(泛化)能力,分别采用 CPP 和 LDA 特征提取后,采用最近邻分类器进行分类.我们给出对数据 SRBCT 和 Brain 实验结果,见图 3.其

中,“LDA 样本空间”表示 LDA 采用“伪逆代替逆”的方法、“LDA 类内散度去奇异”表示 LDA 采用“对类内散布矩阵进行 K-L 展开消除类内散布矩阵奇异”的方法、“LDA 类间散度去奇异”表示 LDA 采用“对类间散布矩阵进行 K-L 展开消除类间散布矩阵奇异”的方法;“CPP 样本空间”表示 CPP 采用采用“伪逆代替逆”的方法、“CPP 类间散度去奇异”表示 CPP 采用“对类间散布矩阵进行 K-L 展开消除类间散布矩阵奇异”的方法.



(*a*) SRBCT



(*b*) Brain

图3　采用最近邻分类器,不同方法在不同维最优子空间的5重交叉验证的分类识别率曲线比较

从图 3 可以看出,总的来看,CPP 的分类正确识别率要明显好于 LDA,而且 CPP 在高维最优子空间的识别率也比较平稳,而 LDA 达到最高识别率后,随着特征维数的增加,识别率明显下降.对于 CPP 来说,采用“对类间散布矩阵进行 K-L 展开消除类间散布矩阵奇异”的鉴别性能又要好于采用“伪逆代替逆”解决矩阵奇异的特征提取的鉴别性能.对 SRBCT 数据,采用“对类间散布矩阵进行 K-L 展开消除类间散布矩阵奇异”的 CPP 在各个不同维的子空间上的正确识别率都要明显高于 LDA,而且在特征子空间维数为 7 时,正确识别率达到了最大值(达到了 98.44%).而 LDA 的最高正确识别是 96.88%.而且在维数大于 7 时,CPP 的正确识别率随着特征子空间维数的增加基本趋于平稳.对数据 Brain,采用 CPP 的整体识别率要高于 LDA,而且 CPP 的正确识别率最高达到了 92.86%(此时的最优子空间维数是 6),而 LDA 的正确识别率最高为 80.95%(此时的最优子空间维数为 4).

综上说明:相比 LDA,CPP 不仅降低了计算复杂度,而且具有更好的鉴别性能,最优子空间的维数不受样本类别数的限制,能找到更高维的最优子空间.

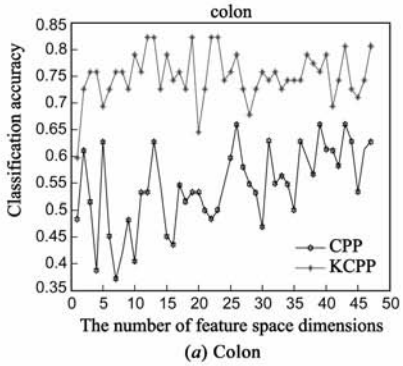
5.2　KCPP 的实验结果

分别采用 CPP 和 KCPP 对数据进行特征提取(其中,KCPP 采用高斯核函数,$\sigma = 10^{-3}$),并采用“伪逆代替逆”的方法解决奇异矩阵问题;然后用最近邻分类器进行分类识别,采用 5 重交叉验证分析其推广能力,循环 10 次求平均正确识别率.图 4 给出数据 Colon、Lymphoma 和 ALLAML 的实验结果.

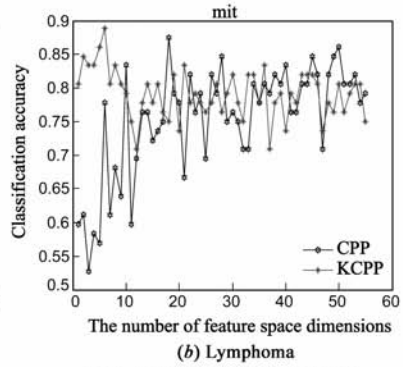
从图 4 可以看出,Colon 数据的 KCPP 识别率在不同维的特征子空间都明显高于 CPP 识别率;Lymphoma 的 KCPP 识别率也基本上高于 CPP;ALLAML 数据在前 9 维特征子空间的 KCPP 识别率要明显高于 CPP.这说明对非线性可分的基因表达数据而言,核类别保留投影也能更好地挖掘数据的类别特征,有效地提取样本的非线性特征.

6　结论

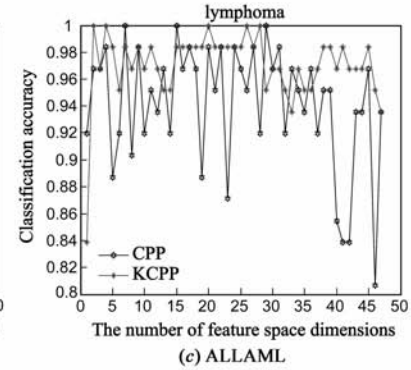
本文提出了一种新的鉴别特征提取方法——类别保留投影(Class Preserving Projection,CPP).通过最小化类内两两样本的平均距离和最大化类间两两样本的平



(*a*) Colon



(*b*) Lymphoma



(*c*) ALLAML

图4　最近邻分类正确识别率

均距离,来寻找最优鉴别特征.与 LDA 最大的不同在于:CPP 类间散布矩阵是由所有样本计算的,而非仅仅是类均值;而且类内散布矩阵和类间散布矩阵是通过矩阵相乘计算得到,而避免了类内协方差矩阵和类间协方差矩阵的计算.相比 LDA,CPP 能保留更多的类间信息,具有更好的鉴别性能;CPP 最优子空间的维数不受样本类别数的限制,能找到更高维的最优子空间;而且 CPP 也大大降低了特征提取的计算复杂度.对基因表达数据的实验,验证了 CPP 的这些优良性能.

通过核技巧,将 CPP 推广到非线性空间,提出核类别保留投影(KCPP),用于解决非线性特征提取问题.对基因表达数据的实验验证了 KCPP 特征提取方法在非线性特征提取上的可行性和有效性.不同形式的核函数其可能的非线性范围很广,故具有很高的灵活性.而如何从众多的核函数中,搜索最优的核函数形式以及其对应的最优参数,是我们今后要做的工作.

参考文献

[1] Ramaswamy S, Golub T R. DNA Microarrays in clinical oncology[J]. Journal of Clinical Oncology, 2002, 20(7): 1932 – 1941.

[2] 孙啸,王晔,赵雨杰,等.一种高密度基因芯片设计的新方法[J].电子学报,2001,29(3):293 – 296.
Sum Xiao, Wang Ye, Zhao Yu jie, et al. A new method of high density gene chip design[J]. Acta Electronica Sinica, 2001, 29(3): 293 – 296. (in Chinese)

[3] 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)

[4] 张艳宁,赵荣椿,梁怡.一种有效的大规模数据的分类方法[J].电子学报,2002,30(10):1533 – 1535.
Zhang Yan-ning, Zhao Rong-chun, Leung Yee. An efficient target recognition method for large scale data[J]. Acta Electronica Sinica, 2002, 30(10): 1533 – 1535. (in Chinese)

[5] 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 Xiao-gang. Cancer subtype recognition and feature selection with gene expression profiles[J]. Acta Electronica Sinica, 2005, 33(4): 651 – 655. (in Chinese)

[6] Hui Gao, James W. Davis. Why direct LDA is not equivalent to LDA[J]. Pattern Recognition, 2006, 39(5): 1002 – 1006.

[7] Sharma Alok, Paliwal Kuldip K. Cancer classification by gradient LDA technique using microarray gene expression data[J]. Data and Knowledge Engineering, 2008, 66(2): 338 – 347.

[8] Paliwal, Kuldip K, Sharma, Alok. Improved direct LDA and its application to DNA microarray gene expression data[J]. Pattern Recognition Letters, 2010, 31(16): 2489 – 2492.

[9] CHENG Zheng-Dong, ZHANG Yu-Jin et al. Study on discriminant matrices of commonly-used fisher discriminant functions [J]. Acta Automatica Sinica, 2010, 36(10): 1361 – 1370.

[10] Q Tian, M Barbero, Z H Gu, S H Lee. Image classification by the Foley-Sammon transform[J]. Opt Eng 1986, 25(7): 834 – 840.

[11] K Liu, Y Cheng, J Yang. A generalized optimal set of discriminant vectors[J]. Pattern Recognition, 1992, 25(7): 731 – 739.

[12] H Yu, J Yang, A direct LDA algorithm for high-dimensional data – with application to face recognition[J]. Pattern Recognition, 2001, 34(10): 2067 – 2070.

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

[14] Pomeroy S L, Tamayo P, Gassenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression[J]. Nature, 2002, 415: 436 – 442.

[15] 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]. Proc Natl Acad Sci, 1999, 96(12): 6745 – 6750.

[16] Alizadeh A A, Eisen M B, Eric Davis R, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling[J]. Nature, 2000, 403: 503 – 511.

[17] Golub T R, Slonim D K, Tamayo P, et al. Molecular classi? cation of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286: 531 – 537.

[18] Gordon G J, Jensen R V, Hsiao L L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma [J]. Cancer Research, 2002, 62: 4963 – 4967.

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

[20] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures[J]. Proc Natl. Acad Sci, 2001, 98(26): 15149 – 15154.

作者简介

王文俊 女,1980 年生于山西平遥,2003 年获西安电子科技大学工学学士学位,并保送攻读本校研究生,2006 年获西安电子科技大学工学硕士学位,2011 年获西安电子科技大学工学博士学位,2006 年至今在西安电子科技大学计算机学院从事教学科研工作,现为讲师,主要研究方向为模式识别、生物信息处理.

E-mail: xidianwwj219@yahoo.com.cn